Genetic variant selection: learning across traits and sites

Laurel Stell[1] and Chiara Sabatti[2]

April 6, 2016

[1] Department of Health Research and Policy, Stanford University, Stanford, CA 94305, U.S.A.
[2] Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305, U.S.A.

Abstract

We consider resequencing studies of associated loci and the problem of prioritizing sequence variants for functional follow-up. Working within the multivariate linear regression framework helps us to account for the joint effects of multiple genes; and adopting a Bayesian approach leads to posterior probabilities that coherently incorporate all information about the variants' function. We describe two novel prior distributions that facilitate learning the role of each variable site by borrowing evidence across phenotypes and across mutations in the same gene. We illustrate their potential advantages with simulations and re-analyzing a dataset of sequencing variants.

1 Introduction

Genome-Wide Association Studies (GWAS) have allowed human geneticists to compile a rather long list of loci where DNA variation appears to be reproducibly associated to phenotypic variability (NHGRI 2015). While these might represent only a subset of the portion of the genome that is important for the traits under study (Manolio *et al.* 2009), there is little doubt that understanding the characteristics and mechanisms of functional variants at these loci is a necessary next step. As resequencing becomes ever more affordable, follow-up investigations of GWAS loci often start with a comprehensive catalogue of their genetic variants in a sample of thousands of individuals, raising the question of how to sort through these results.

Among the many challenges, let us discuss two. First, common variants are often correlated and it is difficult to distinguish their roles without accounting for the broader genetic background of the individuals who carry them. Second, rare variants are present in a small enough portion of the sample that statistical statements become impossible. With this in mind, it has been noted that (a) it is important to account for correlation between variants to obtain useful ranking; (b) we should increasingly be able to take advantage of the information gathered through other studies; and (c) Bayesian models provide a principled approach to guide variant prioritization. To adequately select among variants in the same locus (defined as a genomic region that might encompass multiple genes but that corresponds to the same association signal in a GWAS study), researchers have resorted to model selection approaches (Valdar *et al.* 2012) or approximations of the joint distribution of univariate test statistics (Faye *et al.* 2013; Hormozdiari *et al.* 2014). Prior information on variant annotation has been incorporated in models for eQTL (Veyrieras *et al.* 2008) and more recently for general traits (Pickrell 2014; Kichaev *et al.* 2014; Chung *et al.* 2014),

and annotation programs increasingly attempt to include information on identified genetic loci (Wang *et al.* 2010). Prioritization often relies on Bayes's theorem, and Bayesian methods have received renewed attention in the context of GWAS data analysis (Guan and Stephens 2011; Peltola *et al.* 2012a,b), genomic prediction (Gianola 2013), and the evaluation of heritability (Zhou *et al.* 2013).

In this context, we explore the advantages of a careful specification of the prior distributions on variants, by allowing sharing of information across multiple phenotypes and across neighboring rare variants. We are motivated by the analysis of an exome resequencing study (Service *et al.* 2014) in which individual level data is available for exomic variants at multiple genomic loci that have demonstrated evidence in GWAS of association to lipid traits. By design, the vast majority of measured variants is coding or in UTR, that is, in portions of the genome with high prior probability of harboring functional mutations. Annotation can help distinguish the role of synonymous variants, and conservation scores can be used to predict the effect of nonsynonymous ones; but annotation cannot be used to discount the importance of a large number of non-coding variants that one can expect to occur in a whole genome sequencing dataset. Measures on levels of High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), and triglycerides (TG) are available for the study subjects, and we are interested in capitalizing on the multidimensional nature of the phenotype. Prior analyses of this dataset (Service *et al.* 2014; Bogdan *et al.* in press) have illustrated the importance and the challenges of multivariate linear models, and we explore here the advantages offered by carefully selecting priors for Bayesian models. Abstracting from the specifics of this dataset, we show how hierarchical prior distributions can be adapted to learn about the functionality of a variant by (i) looking across multiple phenotypes and (ii) aggregating the effects of multiple rare variants in the same gene. Since the power of Bayesian methods in borrowing information is well known, it is not surprising that others have explored their application in this context. For example, Yi *et al.* (2011) illustrate the use of priors to model a group effect for multiple rare variants, while Stephens (2013) describes models for the analysis of multiple traits. Our approach, however, is distinct from others in that it strives to achieve all of the following: (1) constructing a multivariate linear model that simultaneously accounts for the contributions of multiple genes and genomic loci; (2) providing inference on variant-specific effects—while linking information across traits and genomic sites; and (3) accounting for the large number of variants tested, effectively enacting a form of multiple comparison adjustment.

This paper is organized as follows. We devote Section 2 to introducing the novel priors in the context of the genetic model, using an approximation of the posterior distribution to illustrate their inferential implications. Section 3 describes the MCMC scheme used to sample the posterior, the setting used for simulations, and the criteria for comparison of methods. Section 4 presents the results of simulation studies highlighting the potential of our proposal, as well as the description of the analysis of the motivating dataset.

2 Prior distributions on genetic variants

One characteristic of a genetic study based on resequencing, as contrasted to genotyping, is that researchers aim to collect a comprehensive catalogue of all genetic variants. This has implications for the statistical models used to analyze the data and the prior assumptions. Let n be the number of subjects in the study and p the number of polymorphic sites assayed. We will use y_i to indicate the phenotypic value of subject i and X_{iv} the genotype of this subject at variant v (typically coded

as minor allele count). The simplest genetic model for a heritable phenotype is of the form

$$y_i = \sum_{k \in \mathcal{T}} G_{ik} + \zeta_i,$$

where ζ_i incapsulate all nongenetic effects and G_{ik} for $k \in \mathcal{T}$ represent the contributions of a set $\mathcal{T}$ of genes that act additively and independently. Without loss of generality and following a standard practice in GWAS, we will assume that the effects of nongenetic determinants of the phenotypes have been regressed out from y_i so that ζ_i can be considered independent 'error' terms. Let us assume that the genetic effects are a linear function of minor allele counts so that

$$y_i = \sum_{v \in \mathcal{V}} \beta_v X_{iv} + \epsilon_i \tag{1}$$

for a set $\mathcal{V}$ of causal variants with ϵ_i *iid* $\mathcal{N}(0, 1/\rho)$. Although this assumption is substantial, it only has the role of simplifying notation. While (1) represents the true genetic architecture of the trait, the membership of $\mathcal{V}$ is unknown in a typical association study, so the relation between the phenotype and genetic variants is expressed as

$$y_i = \sum_{v=1}^{p} \beta_v X_{iv} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \frac{1}{\rho}\right), \tag{2}$$

summing over all variable sites and with the understanding that only an (unknown) subset of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is different from 0. Below we will use the compact matrix notation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Using (2) to describe the relation between traits and genotypes depends heavily on the assumption that a resequencing study assays all variants. In GWAS, on the other hand, causal variants might be untyped, which means their contributions are partially captured by correlated variants and partially included in the error term. It would still be meaningful in that context to use a linear model to link phenotype and genotypes. However, in GWAS, the errors cannot be assumed independent, and the interpretation of the coefficients of $\mathbf{X}$—as well as their prior distribution—is substantially more complicated. We note that mixed effects models can be used to address the first concern (Kang *et al.* 2010).

The parameters in model (2) are $\boldsymbol{\beta}$ and ρ; we now focus on their prior distribution. Following standard practice, we take $\rho \sim \text{Gamma}(\alpha_\rho, \lambda_\rho)$. (See Guan and Stephens (2011) for another approach that specifically targets GWAS and relies on heritability information.) On the vector $\boldsymbol{\beta}$, we want a prior that reflects our model selection goals and our understanding of the genetic architecture. There are several aspects to consider: (a) given the exhaustive nature of the genotyping process, we believe that most of the variants available do not directly influence the trait; (b) it seems reasonable that a variant that influences one trait (so that its effect size is definitely not zero) might also influence other traits; and finally (c) it appears likely that if a rare variant influences the outcome, other nearby rare variants might also have an effect. Our main goal is to describe prior distributions on $\boldsymbol{\beta}$ that incorporate these beliefs. We start by recalling one class of priors that reflect (a) and then move on to generalizations that account for the sharing of information implied by (b) and (c). In what follows, we assume that the allele counts in the column of $\mathbf{X}$ have been standardized to have mean zero and variance one.

2.1 Priors incorporating sparsity

The prior belief that only a fraction of the typed variants has an effect on the phenotype is but one instance of what is a common assumption in high-dimensional statistics, *i.e.* that the

parameter $\boldsymbol{\beta}$ of interest is sparse. To specify a prior on $\boldsymbol{\beta}$ that gives positive probability to vectors with a number of coordinates equal to zero, we rely on a construction by George and McCulloch (1993) and introduce a vector of indicator variables $\mathbf{Z}$ such that $Z_v = 0$ implies $\beta_v = 0$. The Z_v are *iid* Bernoulli with parameter ω, which governs the sparsity of the model and has a Beta(A_ω, B_ω) prior. Let $\boldsymbol{\beta}_Z$ indicate the collection of elements of $\boldsymbol{\beta}$ corresponding to nonzero elements of $\mathbf{Z}$, and let $\mathbf{X}_Z$ be the corresponding columns of $\mathbf{X}$. It has been found useful to assume $(\boldsymbol{\beta}_Z|\mathbf{Z}, \rho, \tau) \sim \mathcal{N}\left(0, \frac{\tau^2}{\rho}\Sigma_Z\right)$, where Σ_Z is a known matrix and $\tau \sim \text{Unif}(\tau_1, \tau_2)$ links the error variance to the size of the $\boldsymbol{\beta}$ coefficients. In the literature, Σ_Z mainly has one of two forms: $I_{|\mathbf{Z}|}$ (the identity matrix of size $|\mathbf{Z}|$, where $|\mathbf{Z}|$ indicates the number of nonzero components of the vector $\mathbf{Z}$) or $n(\mathbf{X}_Z^T\mathbf{X}_Z)^{-1}$, which is referred to as the g-prior (Zellner 1986) (and is a viable choice only when $|\mathbf{Z}| < n$). Various views on the choice of Σ_Z have been put forth (Chipman *et al.* 2001; Heaton and Scott 2010; Guan and Stephens 2011), but the strongest argument for the g-prior is that it provides computational benefits (see below). For either choice of Σ_Z, all of its diagonal entries are equal, resulting in an equal prior variance for each of the β_v. Given the standardization of the columns of $\mathbf{X}$, this implies that the original effect sizes are expected to be larger for rare variants than for common variants, which is reasonable.

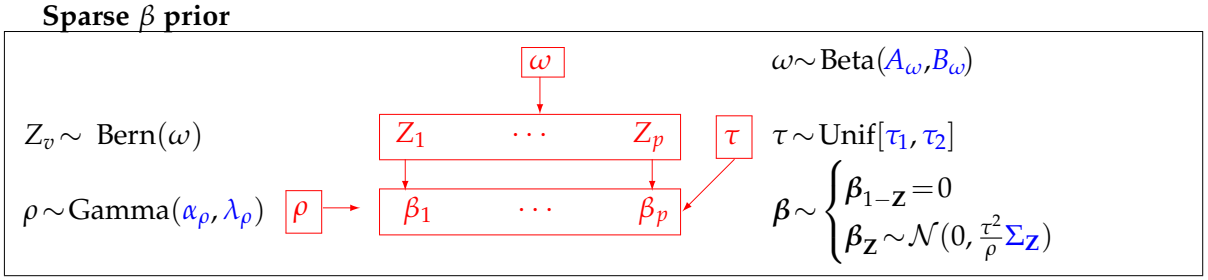

Figure 1: Schematic representation of the *Sparse* prior distribution on $\boldsymbol{\beta}$. Hyperparameters are indicated in blue. The red portion describes all random objects and their dependency structure; unless explicitly indicated, random variables are independent. Boxes identify variables that share one of the distributions depicted in black.

One of the advantages of the prior summarized in Figure (1) is that the derived posterior distribution can be analytically integrated with respect to ω, $\boldsymbol{\beta}$, and ρ. While a MCMC is still needed to fully explore the posterior and carry out inference, we can rely on a collapsed Gibbs sampler that focuses only on τ and the indicator variables $\mathbf{Z}$. This reduces the computation at each iteration and improves its convergence rate (Liu 1994). The prior densities for τ and $\mathbf{Z}$ will be denoted, respectively, as $f_\tau(\tau)$ and $f_Z(\mathbf{Z})$—the latter being easily obtained from the beta-binomial distribution assumed for $|\mathbf{Z}|$. As shown in the appendix, integrating $\boldsymbol{\beta}$ and ρ out gives the marginal posterior density

$$f_{\mathbf{Z},\tau}(\mathbf{Z}, \tau|\mathbf{y}) \propto f_\tau(\tau) f_Z(\mathbf{Z}) \left(\lambda_\rho + \frac{S_Z^2}{2}\right)^{-\left(\frac{n}{2}+\alpha_\rho\right)} \frac{\det(\Omega_Z)^{1/2}}{\tau^{|\mathbf{Z}|}\det(\Sigma_Z)^{1/2}}, \tag{3}$$

where $\Omega_Z^{-1} = \mathbf{X}_Z^T\mathbf{X}_Z + \tau^{-2}\Sigma_Z^{-1}$ and $S_Z^2 = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}_Z\Omega_Z\mathbf{X}_Z^T\mathbf{y}$. Choosing Σ_Z as in the g-prior leads to a simplification of the ratio in (3), thereby avoiding the evaluation of one determinant at each iteration.

Despite the need to evaluate numerically interesting summaries of the posterior of $\mathbf{Z}$, we obtained an approximation (whose derivation and applicability is described in the appendix) to

gain a general understanding of how hyperparameters and data contribute to the final inferential results. Specifically, we focus on the posterior expected value of Z_v, indicator of variant v, conditional on the indicators of all other variants $Z_{[-v]}$. In the case of orthogonal regressors, this expectation can be approximated as

$$\mathbb{E}[Z_v|\mathbf{Z}_{[-v]},\tau,\mathbf{y}]^{-1} \approx 1+\tau\sqrt{n}\frac{B_\omega+p-|\mathbf{Z}_{[-v]}|-1}{A_\omega+|\mathbf{Z}_{[-v]}|}\left(1-\eta_v^2\right)^{n/2}, \tag{4}$$

where $\eta_v = \mathbf{x}_v^T\mathbf{y}/\sqrt{n\mathbf{y}^T\mathbf{y}}$ is approximately the correlation between variant v and the trait. From (4), one gathers that increasing $|\mathbf{Z}_{[-v]}|$, which is the number of variants already used to explain the trait, increases the chance of an additional variant v to be considered relevant. This is a consequence of the fact that the parameter ω, which describes the sparsity of $\boldsymbol{\beta}$ and hence the degree of poligenicity of the trait, is learned from the data (rather than set at a predetermined value). When a large number of variants have been found relevant, the trait is estimated to be highly poligenic and hence it is judged more likely that an additional variant might contribute to its variability. On the other hand, augmenting the total number of genotyped sites p will make it harder for any specific variant v to be judged important; this is adjusting for the look-everywhere effect, an important step in gene mapping studies.

Now that we have introduced this basic framework, we can consider modifications that facilitate learning about the role of a variant across multiple traits and in the context of neighboring sites. We start with the first problem.

2.2 Learning across traits

One of the characteristics of current genetic datasets is the increased availability of multidimensional phenotypes. This is due partly to the automation with which many traits are measured and partly to the increased awareness that precise phenotypic measurements are needed to make progress in our understanding of the underlying biological pathways. Having records of multiple traits in the same dataset allows for cross-pollination of genetic information. On the one hand, if a genetic variant is functional, it can be expected to impact more than one phenotype. On the other hand, even if noise in one phenotype makes it hard to distinguish the predictive power of a causal variant from that of a non-causal neighboring variant, it is much less likely that multiple traits would have noise correlated in such a way that causal and non-causal variants are indistinguishable for all of them. With this in mind, let us generalize the variant selection problem described in the previous section to handle multiple traits.

Extending the notation, let $\mathbf{y}_t$ be the standardized values for trait t, $\boldsymbol{\beta}_t$ the coefficients of $\mathbf{X}$ in the mean $\mathbb{E}[\mathbf{y}_t]$, and $\mathbf{Z}_t$ the corresponding indicator vector. We organize these by column in a $n \times q$ matrix $\mathbf{Y}$, a $p \times q$ matrix $\boldsymbol{\beta}$, and a $p \times q$ matrix $\mathbf{Z}$. Also let $s_v = \sum_t Z_{vt}$ denote the number of traits associated with variant v, let $\boldsymbol{\beta}_{\mathbf{Z}_t}$ be the entries of $\boldsymbol{\beta}_t$ corresponding to entries equal to one in $\mathbf{Z}_t$, and let $\mathbf{X}_{\mathbf{Z}_t}$ be the corresponding columns of $\mathbf{X}$. The data-generating model is $\mathbf{y}_t = \mathbf{X}_{\mathbf{Z}_t}\boldsymbol{\beta}_{\mathbf{Z}_t} + \boldsymbol{\epsilon}_t$ with $\boldsymbol{\epsilon}_t \sim \mathcal{N}\left(0, \frac{1}{\rho_t} I_n\right)$, and the priors on ρ_t are simple extensions of the one used previously: ρ_t *iid* Gamma$(\alpha_\rho, \lambda_\rho)$. Note that this model assumes that, conditionally on the genetic variants that influence them, the traits are independent; specifically, there are no shared environmental effects. This assumption might or might not be appropriate depending on context, but the prior distribution on $\boldsymbol{\beta}$ that we are about to describe can be used also for models that do not rely on this assumption.

We want a prior for $\boldsymbol{\beta}$ that continues to enforce sparsity but that allows learning about the role of a variant across traits. One possibility, first proposed by Jia and Xu (2007), is to introduce

a variant-specific probability of functional effect ν_v, constant across traits and a priori independent with $\nu_v \sim \text{Beta}(A_v, B_v)$, where A_v and B_v can capture annotation information. Following the setup of the previous section, we then take $(Z_{vt}|\nu_v)$ independent Bernoulli(ν_v), set $\beta_{vt} = 0$ whenever $Z_{vt} = 0$, and let $(\boldsymbol{\beta}_{\mathbf{Z}_t}|\mathbf{Z}_t, \rho_t, \tau)$ be independent across t with distribution $\mathcal{N}(0, \frac{\tau^2}{\rho_t}\Sigma_{\mathbf{Z}_t})$. As before, $\tau \sim \text{Unif}(\tau_1, \tau_2)$.

As detailed in the appendix, we can derive an approximation analogous to (4):

$$\mathbb{E}[Z_{vt}|\mathbf{Z}_{[-(vt)]}, \tau, \mathbf{Y}]^{-1} \approx 1 + \tau\sqrt{n}\frac{B_v + q - s_{v,[-t]} - 1}{A_v + s_{v,[-t]}}\left(1 - \eta_{vt}^2\right)^{n/2}, \tag{5}$$

where $\eta_{vt} = \frac{\mathbf{x}_v^T \mathbf{y}_t}{\sqrt{n\mathbf{y}_t^T\mathbf{y}_t}}$ and $s_{v,[-t]} = \sum_{\ell \neq t} Z_{v\ell}$ tallies the number of phenotypes for which the variant v has been judged relevant. This highlights a consequence of the selected prior distribution: as the total number of phenotypes q here has taken the role of p in (4), the role of each variant is judged not in reference to all the other variants but only in comparison to the effect of the same variant across traits. In other words, while there is learning across phenotypes, there is no adjustment for the multiplicity of queried variants. Bottolo *et al.* (2011) previously observed that sparsity of $\mathbf{Z}_t$ could not be controlled by specification of the priors in this approach, and proposed letting Z_{vt} have Bernoulli parameter $\nu_v\omega_t$ with independent priors on each factor.

We propose a different remedy by introducing another layer in the hierarchical priors. Let $\mathbf{W}$ be a vector of indicator variables of length p: if $W_v = 0$, then $\nu_v = 0$; if $W_v = 1$, $\nu_v \sim \text{Beta}(A_v, B_v)$. We take W_v *iid* Bernoulli(ω_W) with $\omega_W \sim \text{Beta}(A_W, B_W)$; the $(Z_{vt}|\nu_v)$ are independent Bernoulli(ν_v), as before. The schematic in Figure 2 summarizes this prior proposal.

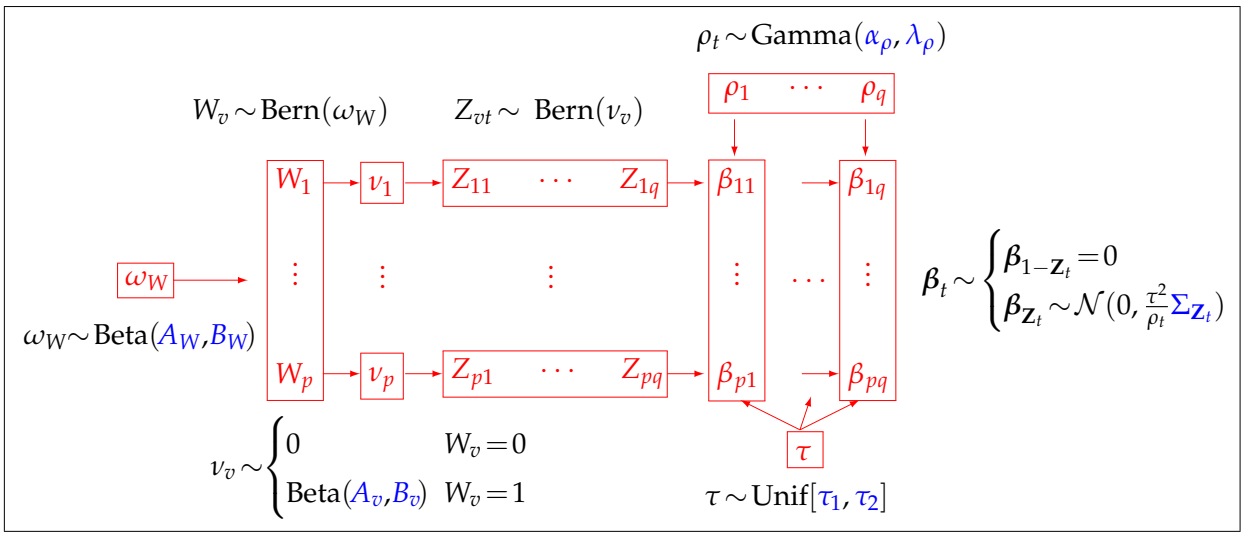

Figure 2: Schematic representation of the *Across Traits* prior distribution on $\boldsymbol{\beta}$. Hyperparameters are indicated in blue. The red portion describes all random objects and their dependency structure; unless explicitly indicated, random variables are independent. Boxes identify variables that share one of the distributions depicted in black.

The existence of a specific parameter ν_v for each site v allows variation in the average number of impacted traits per variant; some variants can be highly pleiotropic, while others are relevant

for one trait only. The sparsity parameter ω_W is once again estimated from the data, allowing for multiplicity adjustment. The introduction of $\mathbf{W}$ effectively specifies a hierarchical prior on $\nu_1, \ldots, \nu_p$; among the many possible ways to accomplish this, the one we adopt emphasizes the role of the sparsity parameter ω_W and is easily interpretable. The appendix presents an indicative approximation of the posterior conditional expected values of W_v similar to (4) and (5). It depends on all phenotypes (enabling learning across traits), but the total number of variants p has again become the leading factor for effective multiplicity correction. We will refer to this prior as learning *Across Traits*. We include the first proposal in some comparison studies, indicating it as the *Unadjusted* approach to emphasize the fact that it does not include an adjustment for multiplicity.

This may be an appropriate point at which to clarify the relation between the prior we are proposing and the traditional investigation of pleiotropy versus coincident linkage. The latter terminology derives from linkage studies, where the nature of the signal is such that the localization of variants with possible pleiotropic effects is possible only up to a certain genomic interval. This interval might contain one variant affecting multiple traits or contain different variants, each affecting a subgroup of the traits. First, it is worth noting that in this paper we are working in the context of association studies, which allow for a much finer resolution than linkage studies. The occurrence of multiple variants affecting multiple traits within the same linkage disequilibrium block is less likely, given that LD blocks are shorter than linkage regions. Secondly, ours is a fixed effects model using sequence data. We are aiming to estimate the specific effect of each variant rather than simply identifying a locus with a random effects model. Our framework, then, automatically considers two options: one variant affecting multiple traits or multiple variants affecting separate traits. The choice between these two alternatives is made on the basis of the posterior probability of the two models. This being said, it is important to recall that if two neighboring variants in LD affect two separate traits, the posterior probabilities of the two alternative models might be similar. The prior we introduce favors pleiotropy in the sense that it recognizes as likely that some variants affect multiple genes, but it does not exclude the alternative explanation, allowing the data to tilt the posterior in either direction. We have investigated this with simulations in the supplementary material.

2.3 Learning across sites

We now consider another form of 'learning from the experience of others' to improve our ability to identify functional variants. We focus on rare variants, which are observed in just a handful of subjects and for which it might be impossible to estimate individual effects. It is reasonable to assume that if one rare variant in a gene has an impact on a trait, other rare variants in the same gene also might be functional; with an appropriate hierarchical prior we might increase our ability to learn the effects of these variants. Of course a similar assumption might also be reasonable for common variants, but given that we observe these in a sufficiently large sample, we aim to estimate their individual effect without convolving their signal with that of others.

The data-generating model is again (2). We define r groups of variants, and we use $\gamma(v)$ to indicate the group to which variant v belongs. Let $\mathbf{G} = (G_1, \ldots, G_r)$ be a vector of indicator variables associated to the groups; we use these to link information from different variants. Specifically, if $G_g = 0$, then the proportion ν_g of causal variants in group g is equal to zero; otherwise, $\nu_g \sim \text{Beta}(A_g, B_g)$ (setting $\nu_g = 1$ for groups comprised of only one variant). The variant-specific indicators Z_v are *iid* Bernoulli with parameter $\nu_{\gamma(v)}$. Similarly to prior specifications, $(G_g|\omega_G)$ are *iid* Bernoulli(ω_G) with $\omega_G \sim \text{Beta}(A_G, B_G)$. This results in the partially exchangeable prior on $\boldsymbol{\beta}$ represented in Figure 3; the parameter ν_g allows sharing information on

functionality across all variants in the same group.

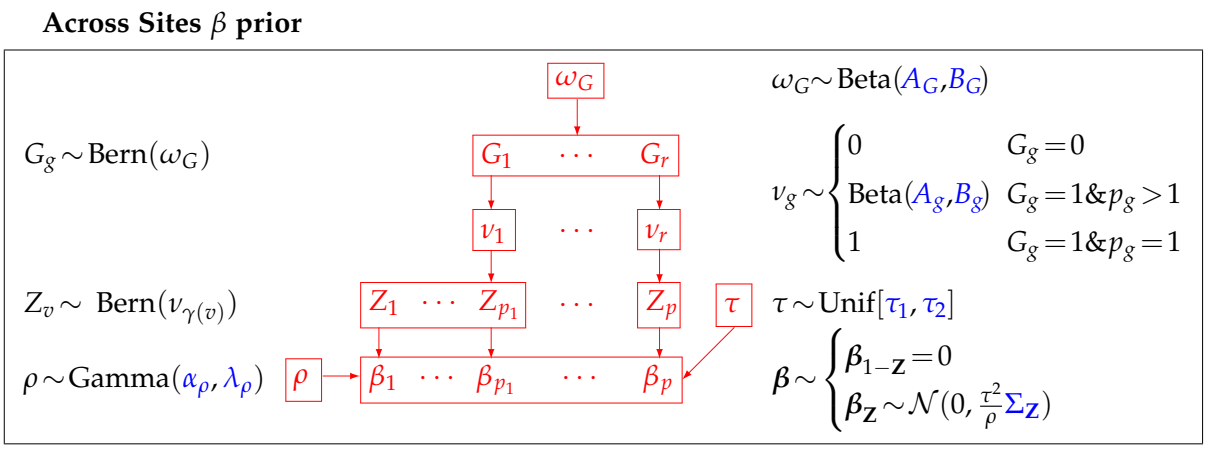

Figure 3: Schematic representation of the *Across Sites* prior distribution on $\boldsymbol{\beta}$. Hyperparameters are indicated in blue. The red portion describes all random objects and their dependency structure; unless explicitly indicated, random variables are independent. Boxes identify variables that share one of the distributions depicted in black.

As described in the appendix, the posterior conditional probability that a variant v belongs to the model depends on the overall number of groups, the number of groups considered relevant, and the number $s_g = \sum_{\gamma(v)=g} Z_v$ of variants in the same group that are deemed functional. The prior distribution in Figure 3, which we refer to as learning *Across Sites*, allows one to achieve an effect similar to that of burden tests, while still providing some variant-specific information (which is in contrast, for example, to the proposal in Yi *et al.* (2011)).

3 Methods

3.1 MCMC sampling

While we have resorted to some analytical approximation for expository convenience, we explore the posterior distribution with MCMC. As previously mentioned, we can focus on sampling τ and all indicator variables. We use a Metropolis-within-Gibbs scheme, with the proposal distributions described below. For τ, the common practice of using a truncated Gaussian works well. The discrete indicator variables pose a greater challenge, even though having integrated out $\boldsymbol{\beta}$ allows us to work with sample space of fixed dimension, eliminating the need for a reversible jump MCMC. When there is only one layer of indicator variables $\mathbf{Z}$, the proposal consists of first choosing with equal probability whether to add or remove a variant and then choosing uniformly among the candidate variants the one for which to propose a change of status. If the prior distribution is described using higher level indicators as well, then proposed changes to both levels must be consistent. If an entry of $\mathbf{W}$ is changed from one to zero, the associated entries of $\mathbf{Z}$ also have to be zeroed; when proposing to change an entry of $\mathbf{W}$ from zero to one, the associated entries of $\mathbf{Z}$ are selected from the prior marginal. Additionally, there are proposal moves that leave $\mathbf{W}$ unchanged but then randomly select one of its nonzero entries and draw a proposal for the associated entries of $\mathbf{Z}$ in a fashion analogous to that described previously. Details of the algorithm are in the supplementary material.

These simple proposal distributions will have trouble in two situations. The most common

is when two or more variants are strongly associated with a phenotype but are also strongly correlated with each other due to LD. Any specific Markov chain will tend to include one of the variants in the model, leaving out the rest. Another problematic situation is when the effects of two variants on a phenotype depend upon each other, so neither variant is likely to enter the model by itself, even if their joint inclusion would be favored by the posterior distribution. Others (Guan and Stephens 2011; Peltola *et al.* 2012a,b) have described proposal distributions that overcome these difficulties and that can be reasonably applied to our setting—even though we do not investigate this in detail, focusing on the description of novel priors.

The average $\overline{\mathbf{Z}}$ of realized values of $\mathbf{Z}$ can be used to summarize the evidence in favor of each variant. Given its practical importance, the basic convergence checks incorporated in our package are based on $\overline{\mathbf{Z}}$. By default, the R code distributed in the package `ptycho` starts four chains from different points, runs each chain for a specified number of MCMC iterations, computes the averages for each chain separately, and then checks the range $\Delta\overline{\mathbf{Z}}$ of these averages. Details on the MCMC can be found in the supplementary material. The algorithm is implemented in the R package `ptycho` (Stell 2015).

3.2 Evaluation of variant selection performance

To investigate the performance of the proposed priors, we apply them to simulated and real data. The posterior distribution can be summarized in multiple ways. One can look for the indicator configuration that receives the highest posterior, for example, or make marginal inference on each variant. Both computational and robustness considerations make it practical to rely on posterior averages $\overline{Z}_{vt}$ for comparisons. In the Bayesian models, then, we consider selecting variant v for trait t if the posterior average $\overline{Z}_{vt}$ is larger than a certain threshold $\xi \in (0,1)$: $\mathcal{S}_t \equiv \{v : \overline{Z}_{vt} > \xi\}$. For benchmarking purposes, we will also analyze the datasets with some non-Bayesian approaches. Specifically we will consider (a) the Lasso (Tibshirani 1996); (b) a set of univariate linear regressions (one for each trait and variant), leading to t-statistics used to test the hypotheses of no association $H_{vt} : \beta_{vt} = 0$ with multiplicity adjustment for the pq hypotheses via the Benjamini-Hochberg (BH) procedure at level α (Benjamini and Hochberg 1995); and (c) multivariate regression including all possible variants, with subsequent tests on the pq null hypotheses for each coefficient incorporating adjustment via the BH procedure at level α. The set of selected variants is equivalent in (a) to the set of estimated nonzero coefficients and in (b) and (c) to the set of variants for which the $H_{vt} : \beta_{vt} = 0$ are rejected. We will refer to these approaches as (a) Lasso, (b) BH marginal, and (c) BH full.

The threshold ξ for Bayesian selection, the penalty of the Lasso, and the level α of BH can all be considered tuning parameters. We will compare the results of different procedures as these are varied (see details in the supplementary material). We base our comparison on an empirical evaluation of power and FDR associated with the different methods. Specifically, for each simulation and each method of analysis, we calculate the proportion of causal variants that are identified and the proportion of selected variants that are in fact false discoveries. The average of these values across multiple simulations is what we refer to as power and FDR in the results. The Bayesian methods also provide an estimate of FDR: if $\overline{Z}_{vt}$ is approximately the probability that variant v is causal for trait t, then the mean of $(1 - \overline{Z}_{vt})$ over the selected variants is the Bayesian False Discovery Rate. We let $\widehat{\text{BFDR}}$ denote this mean and explore how well it approximates (or not) the realized FDP, evaluated across all traits and variants.

3.3 Genotype and phenotype data

Our work has been partially motivated by a resequencing study: Service *et al.* (2014) analyzed targeted exome resequencing data for 17 loci in subjects of Finnish descent (from the 1966 Northern Finland Birth Cohort (NFBC) and the Finland-United States Investigation of NIDDM Genetics study (FUSION)). While the original study considered six quantitative metabolic traits, we focus here on the fasting levels of High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), and triglycerides (TG), transformed and adjusted for confounders as in the initial analyses (see supplementary material). The genotype data was obtained by sequencing the coding regions of 78 genes from 17 loci that had been found by previous GWAS meta-analyses to have a significant association to one of the six traits. In addition, we had access to the first five principal components of genome-wide genotypes. The goal in Service *et al.* (2014) is to identify which variants in these loci are most likely to directly influence the observed variability in the three distinct lipid traits.

Data cleansing and filtering are described in detail in the supplement; here we limit ourselves to note that for the purpose of the simulation study, the collection of variants was pruned to eliminate 550 variants observed only once and to obtain a set of variants with maximal pairwise correlation equal to 0.3 by removing another 558 variants. We excluded singletons from consideration since it would not be possible to make inference on their effect without strong assumptions. Multiple considerations motivated our choice of selecting a subset with only modest correlations: (a) correlated variants make the convergence of MCMC problematic, which might impair our ability to understand the inference derived from the posterior distribution; more importantly, (b) it is very difficult to evaluate and compare the performance of model selection methods in the presence of a high correlation between variants; and finally, (c) statistical methods cannot really choose between highly correlated variants and the selection among these needs to rely on experimental studies. Let us expand on these last two points. Procedures that build a multivariate linear model, such as the Lasso, would select one out of multiple highly correlated variants that have some explanatory power for the response; approaches such as BH marginal would instead tend to select them all; and Bayesian posterior probabilities for each of the variants would reflect the fact that substitutes are available: there will be multiple variants with elevated (if not very high in absolute terms) posterior probability.

It becomes difficult to meaningfully compare FDR and power across these methods, substantially reflecting the fact that the problem is somewhat ill-posed: if multiple highly correlated variants are available, any of them can stand for the others, and it is arbitrary to decide on purely statistical grounds that one belongs to the model while the others do not. Since our goal here is to understand the operating characteristics of the procedures, we found it useful to analyze them in a context where the target is well-identified and the results easily comparable.

After the described pruning, the genetic data used in the simulations contains 5335 subjects and 768 variants. Genotypes were coded with minor allele counts, and missing values (0.04% of genotypes) were imputed using variant average counts for consistency with previous analysis. Observed minor allele frequencies range from 2×10^{-4} to 0.5, with a median of 0.0009 and a mean of 0.02. There are 628 variants with MAF<0.01. Annotation information was obtained as in Service *et al.* (2014), resulting in 61% coding, 34% UTR, and the remainder intragenic. Prior to applying the selection methods, the five genetic principal components along with the intercept were regressed out of both $\mathbf{X}$ and $\mathbf{Y}$, and the columns of both were then standardized.

When studying a real dataset, however, investigators might not be comfortable with such a stringent level of pruning; one might be concerned that variants with important effect are eliminated and that one is essentially reducing the information content of the sample. Indeed,

when analyzing real data, we used a much more comprehensive approach, as described in the case study section.

3.4 Simulation scenarios

We constructed two simulation scenarios: one to simply illustrate the advantages of the proposed priors and the other to investigate their potential in a set-up that models a real genetic investigation.

Illustrative example: orthogonal X.

We set $n = 5000$, $p = 50$, $q = 5$, and $\mathbf{X} = \sqrt{\frac{n-1}{n/p}}(I_p \; I_p \; \cdots \; I_p)^T$ so that $\mathbf{X}^T\mathbf{X} = (n-1)I_p$. In generating $\boldsymbol{\beta}$ and the responses, we want to cover a range of different signal-to-noise ratios. To achieve this, we sample values of the parameters using the distributional assumptions that we described in the specification of the priors. To explore the performance of the *Across Traits* and *Across Sites* models—both when they provide an accurate description of reality as well as when they do not—we use three rules to generate the probability with which each variant is associated to each trait. (a) We sample one sparsity parameter ω for each trait and keep it constant across variants. (b) We sample a probability ν_v for each variant and keep it constant across traits. Finally, (c) we define groups of five variants and sample one probability ν_g of causality for each group of variants and each trait. Rules (a)–(c) are most closely reflected in the prior structure of the basic, *Across Traits* and *Across Sites* models, respectively; and we indicate them as **exchangeable variants**, **pleiotropy**, and **gene effect**. We generate 100 datasets per rule, each with q responses, and analyze them with the described set of approaches. When using Bayesian methods, we rely on non-informative priors (see supplement for details).

Actual genotypes X.

To explore the potential power and FDR in the analysis of the dataset with three lipid traits, we generate artificial phenotypes starting from the available pruned genotypes. We consider a mixture of possible genetic architectures. In the construction of each dataset, (a) one gene is selected uniformly at random for each phenotype and 3–4 of its rare variants are causal (gene effect); (b) 40 distinct common variants are selected uniformly at random and each has probability equal to 0.1 to be causal for each of the phenotypes (thereby substantially representing trait-specific variants); and, finally, (c) 10 additional common variants are selected uniformly at random and each has a probability 0.9 to be causal for each phenotype (pleiotropic effects). This results in traits that are on average determined by 3–4 rare variants in one gene, 4 common variants with effects on one trait only, and 9 common variants with effects across multiple traits. We generated a total of 100 such datasets, as detailed in the supplementary material.

3.5 Data Availability

The sequencing and phenotype data are available on dbGaP. The Northern Finland Birth Cohort 1966 (NFBC1966) study accession number is phs000276.v2.p1. The Finland-United States Investigation of NIDDM Genetics (FUSION) study accession number is phs000867.v1.p1, with the sequencing data in substudy phs000702.v1.p1. In both cases, the sequencing data used in this paper has molecular data type equal to 'Targeted Genome' rather than 'Whole Exome'.

4 Results

4.1 Simulations

Illustrative example.

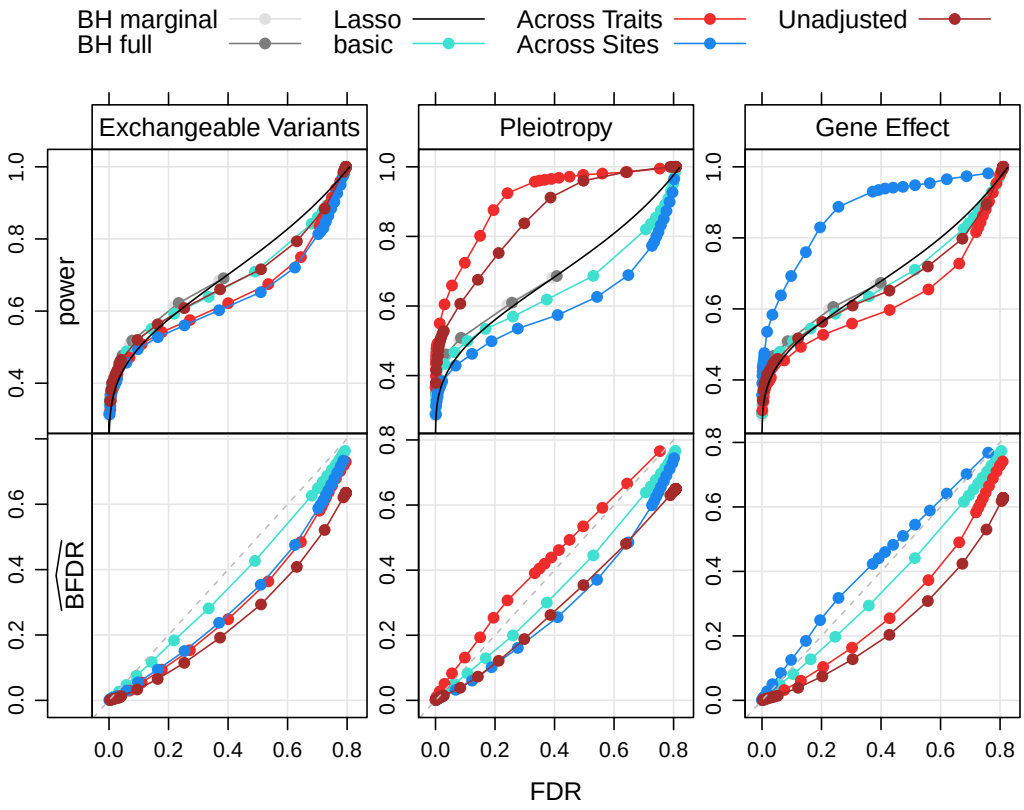

Figure 4: Power (top) and $\widehat{\text{BFDR}}$ (bottom) as a function of empirical FDR in the illustrative example. Each color indicates a different variant selection approach (see legend at the top). Displays in different columns are for different data-generating mechanisms.

Figure 4 showcases the possible advantages of the priors we have described. The plots on the top row compare the empirical FDR and power of the different variant selection methods on the datasets with orthogonal $\mathbf{X}$. Points along the curves are obtained by varying tuning parameters and averaging the resulting FDP and power across 100 simulated datasets. Our setting is such that BH full, BH marginal, Lasso and the basic Bayes model have very similar behaviors: the *Across Traits* and *Unadjusted* models achieve the highest power per FDR in the presence of pleiotropy and the worst power per FDR in the presence of gene effects; in contrast, the *Across Sites* model has maximal power in the presence of gene effects and worse power in the presence of pleiotropy. While it is not surprising that the most effective prior is the one that matches more closely the structural characteristics of the data, it is of note that the loss of power deriving from an incorrect choice of the *Across Traits* or the *Across Sites* model is minimal for FDR values lower than 0.2, which are arguably the range scientists might consider acceptable (see supplementary information for a detail of these values). On the bottom row of Figure 4, we compare the estimated $\widehat{\text{BFDR}}$ with actual FDR for the Bayesian models; here the most serious

mistake is in underestimating FDR, which would lead to an anti-conservative model selection. Once again it can be seen that the best performance is obtained with the prior that matches the data-generating process. Besides this, it is useful to analyze the behavior of the *Unadjusted* approach: its power increase per FDR in the presence of pleiotropy is less pronounced than that of the *Across Traits* model, substantially because the *Unadjusted* approach is too liberal, with a $\widehat{\text{BFDR}}$ which is significantly underestimated. This is in agreement with the lack of adjustment for multiplicity indicated by (5). Results for alternate hyperparameters are in the supplement.

Generating phenotypes from actual genotype data.

Figure 5 shows the performance of the variant selection methods in the analysis of traits generated from actual genotype data, further emphasizing the potential gains associated with the proposed strategies. For given FDR, both the *Across Traits* and *Across Sites* priors lead to an in-

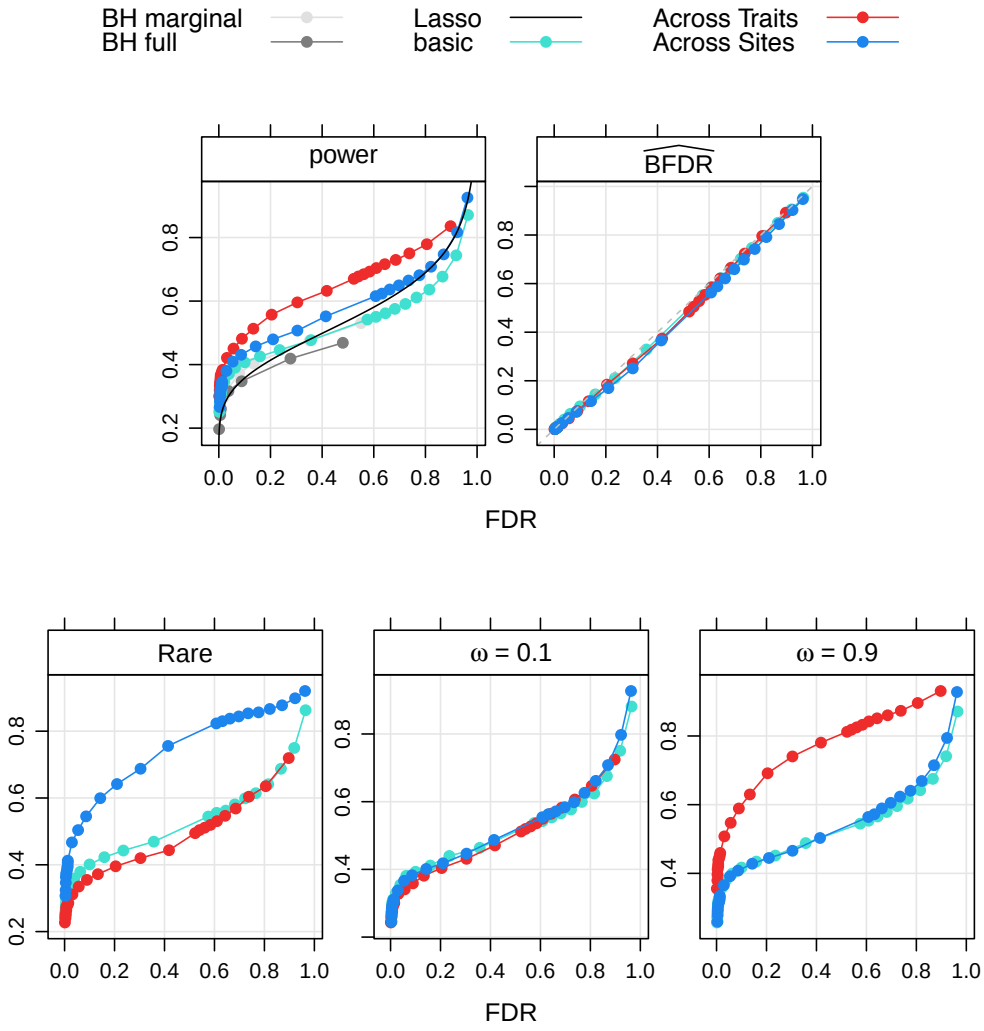

Figure 5: In the top portion, power and $\widehat{\text{BFDR}}$ as a function of empirical FDR in the simulation from actual genotype data. In the lower panels, power is calculated separately for the different types of variants: rare, common with trait-specific effects, and common with pleiotropic effects.

Table 1: FDR and power for specific choices of selection parameters applied to simulated traits with actual genotype data.

Variant selection criteria	**FDR**	**power**
BH full with $\alpha = 0.05$	0.037	0.32
BH marginal with $\alpha = 0.05$	0.079	0.37
Lasso min error λ	0.810	0.63
Lasso 1-se λ	0.046	0.26
basic $\widehat{\text{BFDR}} \leq 0.05$	0.039	0.37
Across Traits $\widehat{\text{BFDR}} \leq 0.05$	0.054	0.45
Across Sites $\widehat{\text{BFDR}} \leq 0.05$	0.051	0.40

crease in power over the other methods. This is due to the fact that phenotypes are generated assuming both pleiotropy and contributions from multiple rare variants in the same gene (gene effects). In the lower portion of Figure 5, we separate the power to recover rare variants with gene effects from that for trait-specific common variants ($\omega = 0.1$) and from that for common variants with pleiotropic effects ($\omega = 0.9$). As expected, the gains of *Across Traits* and *Across Sites* are for the portion of genetic architecture that is accurately reflected in these priors. The estimates $\widehat{\text{BFDR}}$ are accurate, indicating that all three Bayesian priors correctly learned τ and the probabilities of function.

Finally, while we have relied on ROC-like curves to compare different approaches as the value of their tuning parameters vary, it is useful to focus on the operating characteristics of the standard ways of selecting the tuning parameters. By convention, the target FDR for BH is usually 0.05. For Lasso selection, the function `cv.glmnet` provides two choices for λ: minimizing the cross-validation error and using the one-standard error rule. In Bayesian approaches, one can select variants so that global $\widehat{\text{BFDR}} \leq 0.05$. Table 1 compares FDR and power for these selection parameters; the Bayesian methods appear to control the target FDR and arguably result in better power. Analogous summaries for other decision rules are included in the supplementary material; here we simply remark that including variants such that $\widehat{\text{BFDR}} \leq 0.05$ in this dataset was practically equivalent to selecting variants with posterior probability larger than 0.7. We will capitalize on this observation for the real data analysis.

The supplement details the results of another set of simulations along the lines of a traditional investigation of pleiotropy versus coincident linkage; we give a very brief summary here. In the case of separate causal variants, the *Across Traits* prior may have a slight loss of power but is still much better than BH with p-values from the full model. In the case of pleiotropy, however, the *Across Traits* prior clearly has greater power per FDR.

4.2 Case Study: the influence of 17 genomic loci on lipid traits

We now turn to the analysis of the three lipid traits in the Finnish dataset. While resequencing data comes from 17 loci identified via GWAS, prior evidence of association is available only between some of these loci and some traits. In particular, four loci have no documented association with any of the three lipid traits we study; we include variants from these loci in the analysis as negative controls. (This is different from the work in Service *et al.* (2014), which examines only variants in loci specifically associated with each trait.)

Table 2: Summary of selections for BH with $\alpha = 0.05$, Lasso with λ chosen by `cv.glmnet`, and Bayesian approaches with $\xi = 0.7$. The columns labeled R and V^* give, respectively, the number of variants selected across the entire study and in the four loci with no prior evidence of association to any of the lipid traits analyzed (CRY2, G6PC2, MTNR1B, and PANK1). $\widehat{\text{BFDR}}$ reports, for Bayesian methods, the Bayesian FDR computed separately for each trait.

	HDL			LDL			TG		
Variant Selection	R	V^*	$\widehat{\text{BFDR}}$	R	V^*	$\widehat{\text{BFDR}}$	R	V^*	$\widehat{\text{BFDR}}$
BH full p-values	13	0		3	0		5	0	
BH marginal p-values	22	0		6	0		10	0	
Lasso min error λ	134	12		40	4		80	6	
Lasso 1-se λ	16	0		0	0		4	0	
basic	21	0	0.046	4	0	0.012	8	0	0.011
Across Traits	19	0	0.084	8	0	0.059	9	0	0.029
Across Sites	25	0	0.064	5	0	0.063	8	0	0.007

Service *et al.* (2014) relied on univariate regression to test the association between each trait and each variant with MAF>0.01 and on burden tests to evaluate the role of nonsynonymous rare variants. Bogdan *et al.* (in press) re-analyzed the data relative to HDL with a set of model selection approaches, including the novel methodology SLOPE; to facilitate comparison with their results, we add SLOPE to the analysis methods considered so far. Groups for the *Across Sites* model were defined so as to mimic the burden tests in Service *et al.* (2014), which means a group with more than one variant contains all nonsynonymous variants with MAF<0.01 in the same gene.

We start by analyzing the pruned subset of variants used in the simulation studies and postpone a more exhaustive search, noting again that this allows for a more straightforward comparison of the variants selected by different methodologies. Table 2 compares the number of variants selected by various methods with specified tuning parameters. The column labeled V^* shows the number of selected variants that are in a locus lacking any prior evidence of association to lipid traits. The Lasso with λ chosen to minimize cross-validated prediction error clearly results in far too many selections, so we discard this approach for the remaining results. For Bayesian approaches, the threshold $\xi = 0.7$ results in average $\widehat{\text{BFDR}}$ approximately controlled at the 0.05 level.

Figure 6 illustrates the model selection results for HDL. (Analogous displays for the other two phenotypes are available in the supplementary material, as well as a table detailing differences in selections between approaches.) Each display corresponds to a locus, with turquoise shading (rather than orange) used to indicate prior evidence of association to HDL. Variants are arranged according to their genomic positions in the loci, and the values of their estimated coefficients are plotted on the y-axis; with the exception of marginal BH, we display only nonzero coefficients. When available, a vertical black line indicates the position of the SNPs originally used to select the locus ('Array SNP').

There is substantial overlap among the results of various methods. Model selection approaches seem to generally agree with the findings in Service *et al.* (2014) (with Lasso 1-se the most discrepant, missing a number of the associations identified in Service *et al.* (2014); see supplementary material). Still, we can point to some significant differences. With the *Across Traits* approach we select two variants in two loci where no other method identifies any signal: in

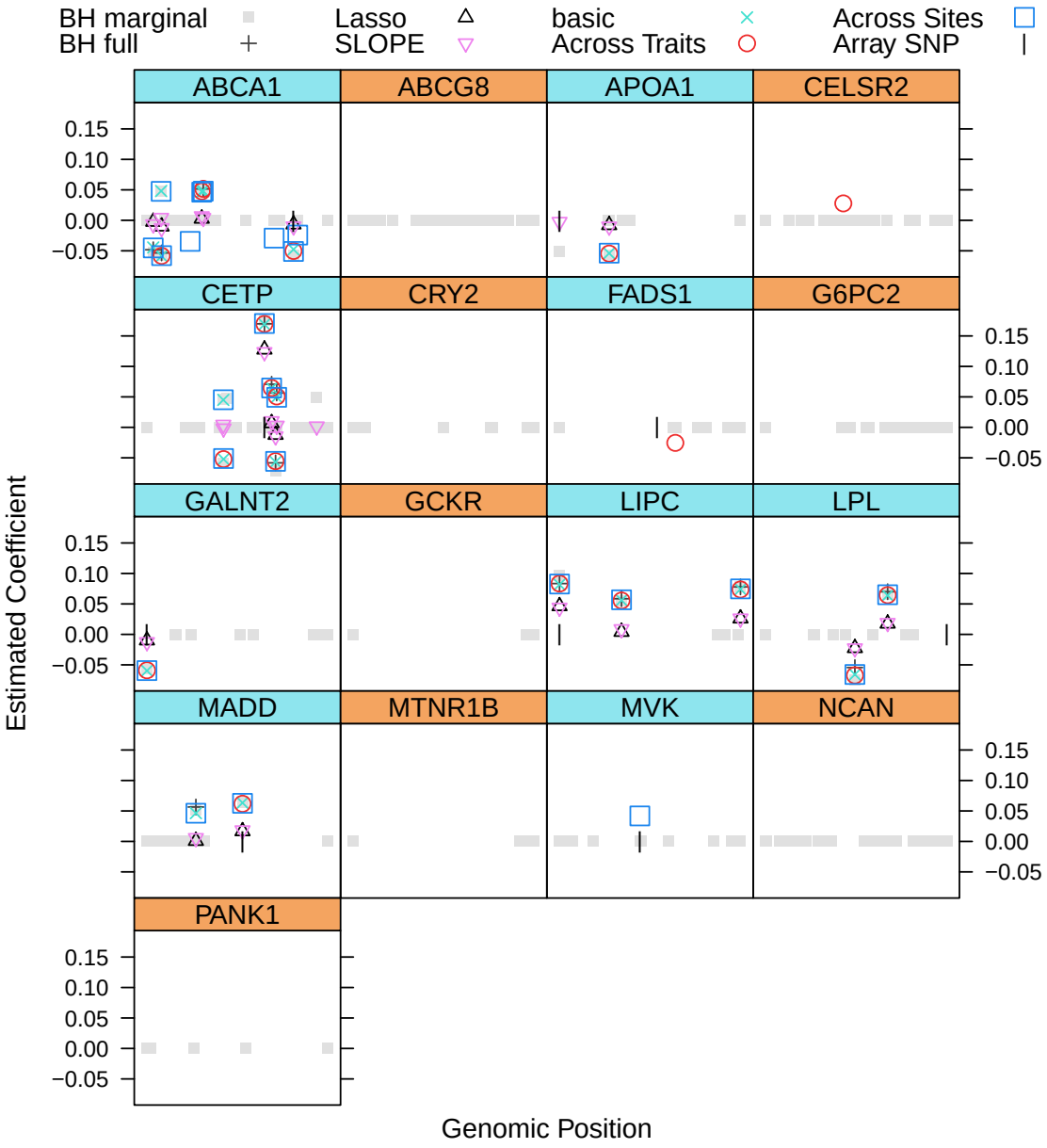

Figure 6: Estimated variant effects on HDL. Each panel corresponds to a locus, the x-axis indicates the variant's genomic position and the y-axis its regression coefficient (with the exception of BH marginal, only nonzero coefficients are represented). The color code of the panel titles indicates the presence/absence of prior evidence of association between the locus and HDL (turquoise/orange, respectively). Model selection methods are distinguished using plotting symbols, as indicated in the legend at the top.

CELSR2 and FADS1. These two loci have prior evidence of association to LDL and to all three lipid traits, respectively, and the *Across Traits* approach identifies pleiotropic effects. In contrast, the *Across Traits* approach does not select four very rare (MAF<0.001) variants considered relevant by more than one alternative method. While we do not know where the truth lies at this point, it is very hard to evaluate the effect of a rare variant on purely statistical grounds, and the outcome of the *Across Traits* model might well be the more reliable.

The *Across Sites* approach identifies four variants that other approaches overlook. Three are rare variants in ABCA1: two missense rare (0.01>MAF>0.001) and one nonsense very rare (MAF 0.00016); their discovery is facilitated by the fact that they are included in a group with multiple other significant variants. The fourth is a common variant in the MVK locus, for which there is prior evidence of association to HDL. Other approaches do not recover this simply because the signal in the locus is, in the dataset analyzed here, barely below the detection threshold; *Across Sites* has a slight advantage over the other Bayesian methods because grouping reduces the number of comparisons to account for. We note that ABCA1 is a gene in which rare variants were found to have a role by the burden test analysis in Service *et al.* (2014).

We have relied on the pruned dataset since low correlation across variants greatly facilitates the comparison of the selection results by different methods. However, mindful of the concerns of scientists unwilling to pre-screen the genotypic information, we have also carried out a more comprehensive analysis of this dataset, showcasing that this is indeed an option available to researchers. Details for these analyses are in the supplement, but we summarize them here. First, we have compared the results of four different levels of pruning (correlation less than 0.3, 0.5, 0.7 and 0.9). We have found that very few values of $\overline{Z}_{vt}$ change by more than 0.05 when different levels of pruning are used and that less stringent levels of pruning do not lead to substantially more findings—unlike when applying BH to marginal p-values. In fact, there is a greater tendency for variants to drop out of the selection set as correlated variants are added to $\mathbf{X}$. Second, to completely eliminate pruning, we analyzed all variants in *one* locus along the lines of Servin and Stephens (2007); Hormozdiari *et al.* (2014); Kichaev *et al.* (2014); Chen *et al.* (2015), using the basic prior and assuming that the number of significant regressors $|\mathbf{Z}|$ is no greater than five or six (depending on the total number of possible variants in the locus). We have restricted our attention to two loci only, those that showed stronger evidence of influencing HDL via multiple variants. The supplementary material offers a precise comparison of results, but it suffices here to note that the set $\mathbf{Z}$ of variants with the largest posterior density is equal to the variants selected among the original pruned set (correlation less than 0.3) by the basic prior for one locus and, for the other locus, the two sets substantially overlap.

5 Discussion

As the genetics community devotes increasing effort to follow up GWAS hits with resequencing, a number of suggestions on how to prioritize variants have emerged. In truth, while dealing with the same broad scientific goals, many of these contributions address different aspects of the problem and therefore should be seen as complementary rather than alternatives; taken together they provide scientists with useful suggestions. Annotation information has been shown to be quite useful when the set of variants under consideration is sufficiently diverse. It is important to account for the correlation across variants to avoid paying attention to SNPs that are only 'guilty by association.' Bayes's theorem and Bayesian statistics are a natural way of dealing with the decision of which variants to pursue. In this context, others have studied (a) how to choose priors that incorporate annotation information, tuning their parameters with

available datasets; (b) how to approximate posterior distributions of variant effects; (c) how to sample from the posterior distribution using efficient MCMC or variational schemes; and (d) how to efficiently evaluate posterior probabilities for a set of variants. Here we focus on another aspect of prior selection: describing how partial exchangeability assumptions can be used to borrow information across traits and neighboring sites, while maintaining an effective control for multiplicity across variants and fitting multivariate regression models that estimate the specific contribution of each associated site, while accounting for others. We briefly refer to some of the most direct antecedents of our proposed priors to underscore relevant differences.

Yi *et al.* (2011) proposed the use of hierarchical priors to capture effects of rare variants through groups, similar to the *Across Sites* model. However, their proposal does not incorporate sparsity considerations, resulting in the estimate of a nonzero effect for each variant and each group and therefore not engaging in model selection. Quintana *et al.* (2011, 2012) took an additional step towards the *Across Sites* model by incorporating sparsity via the indicator variable $\mathbf{Z}$. They considered only rare variants, used the same effect size for all rare variants in a genomic region, used the MLE for the effect sizes rather than integrating them out, and, most importantly, controlled sparsity by using $A_g = 1$ and $B_g = p_g$ in the prior for ν_g rather than introducing another layer of indicator variables in the hierarchical prior—all of which means their approach has less flexibility and less learning.

The *Across Sites* prior also echoes the proposal of Zhou *et al.* (2010) who suggested the use of group penalization in Lasso to estimate multivariate sparse models while encouraging coordinated selection of rare variants in the same gene. This computationally appealing approach has not become as popular in genetics as in many other fields, possibly because of the difficulties connected with the selection of its tuning parameters when model selection is the goal. Cross validation is often used to determine the appropriate level of penalization; while this works well for prediction purposes, its performance is less than satisfactory in identifying variants with truly nonzero coefficients (as illustrated by our case study). Alexander and Lange (2011), Valdar *et al.* (2012), and, most recently, Sabourin *et al.* (2015) explore coupling resampling techniques with Lasso penalization to improve model selection. This not only increases computational costs but also greatly reduces the initial simplicity of the model. As documented in Bogdan *et al.* (in press), identifying a single λ value that performs FDR control is challenging; Yi *et al.* (2015) investigates this task in the context of GWAS and provides guidelines. The final model selection of these machine learning approaches uses complex rules; in contrast, the Bayesian models we described are based on easy to interpret parameters.

The use of hierarchical Bayesian methods has ample precedents in eQTL studies, where they have been used to correct for multiplicity (Kendziorski *et al.* 2006) and to increase power of detecting variants affecting multiple traits. In our presentation of the *Unadjusted* and *Across Traits* approaches, we referred to methods proposed by Jia and Xu (2007) and Bottolo *et al.* (2011). More recent work (Flutre *et al.* 2013; Li *et al.* 2013) has focused on the identification of local (cis) effects across tissue, and considered models with only one functional variant. The recent contribution by Chung *et al.* (2014)—which appeared while this work was in preparation—underscores as we do the importance of learning both across sites and across traits to prioritize variants. These authors, however, work with p-values from GWAS studies, rather than actual resequencing data.

Having clarified the scope of our contribution, we want to briefly mention how it could be extended and combined with suggestions by others. First, let us point out that while in the simulations and in the analytical approximations we assumed $n > p$, this restriction is by no means necessary to the Bayesian model we describe. On the contrary, the priors we propose—by learning sparsity and giving positive probabilities to configurations with some $\beta_v = 0$—are well-suited to the case $n < p$. The real challenge in dealing with GWAS-type data would be from

a computational standpoint: increased mixing for MCMC as described in Xu *et al.* (2014) and Guan and Stephens (2011) or other algorithmic improvements (Carbonetto and Stephens 2012; Hormozdiari *et al.* 2014) would make our approach more widely applicable.

Another extension that is easily achieved is the combination of the *Across Traits* and *Across Sites* priors. Most immediately, the group indicators $\mathbf{G}$ in Figure 3 can be made trait-specific and linked across phenotypes with the same approach used to link the $\mathbf{Z}_t$ in Figure 2.

It is certainly possible to combine the partial exchangeability aspects of our models with a prior that incorporates annotation information. Refer, for example, to the *Across Traits* prior in Figure 2. Currently, the distribution on $W_1, \ldots, W_p$, indicators of functionality of the variants, is a beta-binomial. However, it is trivial to change it to a mixture of independent logits, with the linear model component including an intercept effect—which would capture the overall sparsity—and a linear combination of annotation indicators (Veyrieras *et al.* 2008; Kichaev *et al.* 2014; Pickrell 2014).

Since our focus has been on specification of the prior, we have not paid much attention to the data-generating model, which could certainly be improved. Specifically, we want to underscore the fact that using a mixed-model approach might be advisable to account for population structure (Kang *et al.* 2010) and when analyzing many phenotypes whose quantitative value might be influenced by confounders (Zhou and Stephens 2014) or simply by genetic variants not included in the model.

In conclusion, we want to emphasize the increasing importance in human genetics of models that account for pleiotropy. 'Big data' in genetics has often been equated with the abundance of sequences, and these certainly pose a number of management and interpretation challenges. Our increased acquisition capacity will also result, however, in the collection of a large number of phenotypes; gene expression, MRI scans, and mass spectrometry are just some examples of the large-scale phenotyping efforts underway. Now that DNA variation has been extensively described, annotating this appears as a fundamental challenge; the rich phenotypic collections increasingly available have a major role to play. After all, what better way of establishing if a variant has some functional impact than looking for its association with any trait available? Bayesian models that allow one to estimate the probability with which a variant has functional effects across phenotypes are likely to be very useful. In this paper, we have described a first step in this direction.

Acknowledgements

L. Stell and C. Sabatti were partially supported by NIH grants HG006695, HL113315, MH105578, and MH095454. We thank the authors of Service *et al.* (2014) for letting us use their data during the completion of dbGaP release. Finally our gratitude to H. Tombropoulos for unwavering editorial assistance.

References

D. H. Alexander and K. Lange, 2011, Stability selection for genome-wide association. Genet. Epidemiol., 35(7): 722–728.

Y. Benjamini and Y. Hochberg, 1995, Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol., 57: 289–300.

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candes, in press, SLOPE—adaptive variable selection via convex optimization. Ann. Appl. Stat.

L. Bottolo, E. Petretto, S. Blankenberg, F. Cambien, S. A. Cook, *et al.*, 2011, Bayesian detection of expression quantitative trait loci hot spots. Genetics, 189(4): 1449–1459. doi: 10.1534/genetics.111.131425.

P. Carbonetto and M. Stephens, 2012, Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal., 7(1): 73–108. doi: 10.1214/12-BA703.

W. Chen, B. R. Larrabee, I. G. Ovsyannikova, R. B. Kennedy, I. H. Haralambieva, *et al.*, 2015, Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. Genetics, 200(3): 719–736. doi: 10.1534/genetics.115.176107.

H. Chipman, E. I. George, and R. E. McCulloch, 2001, The practical implementation of Bayesian model selection, pp. 65–134, in *Model Selection*, edited by P. Lahiri, volume 38 of *Institute of Mathematical Statistics Lecture Notes Monograph*. Institute of Mathematical Statistics.

D. Chung, C. Yang, C. Li, J. Gelernter, and H. Zhao, 2014, GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLOS Genet., 10(11): e1004787. doi: 10.1371/journal.pgen.1004787.

L. L. Faye, M. J. Machiela, P. Kraft, S. B. Bull, and L. Sun, 2013, Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. PLOS Genet., 9(8): e1003609. doi: 10.1371/journal.pgen.1003609.

T. Flutre, X. Wen, J. Pritchard, and M. Stephens, 2013, A statistical framework for joint eQTL analysis in multiple tissues. PLOS Genet., 9(5): e1003486. doi: 10.1371/journal.pgen.1003486.

E. I. George and R. E. McCulloch, 1993, Variable selection via Gibbs sampling. JASA, 88(423): 881–889.

D. Gianola, 2013, Priors in whole-genome regression: The Bayesian alphabet returns. Genetics, 194(3): 573–596. doi: 10.1534/genetics.113.151753.

Y. Guan and M. Stephens, 2011, Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. Ann. Appl. Stat., 5(3): 1780–1815.

David A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag, 2008.

M. J. Heaton and J. G. Scott, 2010, Bayesian computation and the linear model, pp. 527–545, in *Frontiers of Statistical Decision Making and Bayesian Analysis*, edited by M. Chen, D. K. Dey, P. Müller, D. Sun, and K. Ye. Springer.

F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, and E. Eskin, 2014, Identifying causal variants at loci with multiple signals of association. Genetics, 198(2): 497–508. doi: 10.1534/genetics.114.167908.

Z. Jia and S. Xu, 2007, Mapping quantitative trait loci for expression abundance. Genetics, 176 (1): 611–623. doi: 10.1534/genetics.106.065599.

H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, *et al.*, 2010, Variance component model to account for sample structure in genome-wide association studies. Nat. Genet., 42(4): 348–354.

C. M. Kendziorski, M. Chen, M. Yuan, H. Lan, and A. D. Attie, 2006, Statistical methods for expression quantitative trait loci (eQTL) mapping. Biometrics, 62(1): 19–27. doi: 10.1111/j.1541-0420.2005.00437.x.

G. Kichaev, W. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, *et al.*, 2014, Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLOS Genet., 10(10): e1004722. doi: 10.1371/journal.pgen.1004722.

G. Li, A. A. Shabalin, I. Rusyn, F. A. Wright, and A. B. Nobel, 2013, An empirical Bayes approach for multiple tissue eQTL analysis. arXiv:1311.2948.

J. S. Liu, 1994, The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. JASA, 89(427): 958–966.

G. Malsiner-Walli and H. Wagner, 2011, Comparing spike and slab priors for Bayesian variable selection. Austrian J. of Stat., 40(4): 241–264.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, *et al.*, 2009, Finding the missing heritability of complex diseases. Nature, 461(7265): 747–753.

NHGRI. *A Catalog of Published Genome-Wide Association Studies*, 2015. `http://www.genome.gov/gwastudies/` [Accessed: 2015].

T. Peltola, P. Marttinen, A. Jula, V. Salomaa, M. Perola, *et al.*, 2012a, Bayesian variable selection in searching for additive and dominant effects in genome-wide data. PLOS ONE, 7: e29115. doi: 10.1371/journal.pone.0029115.

T. Peltola, P. Marttinen, and A. Vehtari, 2012b, Finite adaptation and multistep moves in the Metropolis-Hastings algorithm for variable selection in genome-wide data. PLOS ONE, 7: e49445. doi: 10.1371/journal.pone.0049445.

J. K. Pickrell, 2014, Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet., 94(4): 559–573.

M.Ã. Quintana, J.L̃. Berstein, D.C̃. Thomas, and D.Ṽ. Conti, 2011, Incorporating model uncertainty in detecting rare variants: The Bayesian risk index. Genetic Epidemiology, 35(7): 638–649. doi: 10.1002/gepi.20613.

M.Ã. Quintana, F.R̃. Schumacher, Casey G. J.L̃. Bernstein, L. Li, *et al.*, 2012, Incorporating prior biologic information for high-dimensional rare variant association studies. Hum. Hered., 74: 184–195. doi: 10.1159/000346021.

J. Sabourin, A. B. Nobel, and W. Valdar, 2015, Fine-mapping additive and dominant SNP effects using group-LASSO and fractional resample model averaging. Genet. Epidemiol., 39(2): 77–88.

S. K. Service, T. M. Teslovich, C. Fuchsberger, V. Ramensky, P. Yajnik, *et al.*, 2014, Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. PLOS Genet., 10(1): e1004147. doi: 10.1371/journal.pgen.1004147.

Bertrand Servin and Matthew Stephens, 2007, Imputation-based analysis of association studies: Candidate regions and quantitative traits. PLOS Genet., 3(7): e114. doi: 10.1371/journal.pgen.0030114.

L. Stell. `ptycho`: Bayesian Variable Selection with Hierarchical Priors, 2015. `https://cran.r-project.org/web/packages/ptycho/` [Accessed: 2015].

M. Stephens, 2013, A unified framework for association analysis with multiple related phenotypes. PLOS ONE, 8(7): e65245. doi: 10.1371/journal.pone.0065245.

R. Tibshirani, 1996, Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol., 58: 267–288.

W. Valdar, J. Sabourin, A. Nobel, and C. C. Holmes, 2012, Reprioritizing genetic associations in hit regions using LASSO-based resample model averaging. Genet. Epidemiol., 36(5): 451–462.

J. Veyrieras, S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, *et al.*, 2008, High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLOS Genet., 4: e1000214. doi: 10.1371/journal.pgen.1000214.

K. Wang, M. Li, and H. Hakonarson, 2010, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res., 38(16): e164.

L. Xu, R. Craiu, L. Sun, and A. Paterson, 2014, Parameter expanded algorithms for Bayesian latent variable modelling of genetic pleiotropy data. J. Comp. Graph. Stat. URL `http://www.tandfonline.com/doi/abs/10.1080/10618600.2014.988337#`[Accessed: 2015].

H. Yi, P. Breheny, N. Imam, Y. Liu, and I. Hoeschele, 2015, Penalized multimarker vs. single-marker regression methods for genome-wide association studies of quantitative traits. Genetics, 199(1): 205–222. doi: 10.1534/genetics.114.167817.

N. Yi, N. Liu, D. Zhi, and J. Li, 2011, Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. PLOS Genet., 7: e1002382.

A. Zellner, 1986, On assessing prior distributions and Bayesian regression analysis with g-prior distributions, pp. 233–243, in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, edited by P. K. Goel and A. Zellner. North-Holland/Elsevier.

H. Zhou, M. E. Sehl, J. S. Sinsheimer, and K. Lange, 2010, Association screening of common and rare genetic variants by penalized regression. Bioinformatics, 26(19): 2375–2382.

X. Zhou and M. Stephens, 2014, Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods, 11(4): 407–409.

X. Zhou, P. Carbonetto, and M. Stephens, 2013, Polygenic modeling with Bayesian sparse linear mixed models. PLOS Genet., 9(2): e1003264.

A Mathematical details for the *basic* prior

First we integrate $\boldsymbol{\beta}$ out of the posterior distribution as in (Chen *et al.* 2015). Since $\mathbf{y} = \mathbf{X}_Z\boldsymbol{\beta}_Z + \boldsymbol{\epsilon}$ with $(\boldsymbol{\beta}_Z|\mathbf{Z},\rho,\tau) \sim \mathcal{N}\left(0, \frac{\tau^2}{\rho}\Sigma_Z\right)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \frac{1}{\rho}I_n)$,

$$(\mathbf{y}|\mathbf{Z},\rho,\tau) \sim \mathcal{N}\left(0, \frac{1}{\rho}I_n + \frac{\tau^2}{\rho}\mathbf{X}_Z\Sigma_Z\mathbf{X}_Z^T\right). \tag{6}$$

While the likelihood can be written directly from this, a few manipulations give it in a more convenient form. Sylvester's determinant theorem (Harville 2008, p. 420) implies

$$\det(I_n + \tau^2\mathbf{X}_Z\Sigma_Z\mathbf{X}_Z^T) = \det(I_{|\mathbf{Z}|} + \tau^2\Sigma_Z\mathbf{X}_Z^T\mathbf{X}_Z) = \tau^{2|\mathbf{Z}|}\frac{\det(\Sigma_Z)}{\det(\Omega_Z)},$$

and a generalization of the Woodbury matrix identity (Harville 2008, p. 428) implies

$$(I_n + \tau^2\mathbf{X}_Z\Sigma_Z\mathbf{X}_Z^T)^{-1} = I_n - \mathbf{X}_Z(\tau^{-2}\Sigma_Z^{-1} + \mathbf{X}_Z^T\mathbf{X}_Z)^{-1}\mathbf{X}_Z^T = I_n - \mathbf{X}_Z\Omega_Z\mathbf{X}_Z^T.$$

Consequently, $\Pr(\mathbf{y}|\mathbf{Z},\rho,\tau) = \left(\frac{\rho}{2\pi}\right)^{n/2}\frac{\det(\Omega_Z)^{1/2}}{\tau^{|\mathbf{Z}|}\det(\Sigma_Z)^{1/2}}e^{-\rho S_Z^2/2}$. For the null model $\mathbf{Z} = 0$, (6) shows that the covariance matrix is $\rho^{-1}I_n$, so in this case $S_Z^2 = \mathbf{y}^T\mathbf{y}$ and the ratio of determinants is set equal to one.

Next multiply $\Pr(\mathbf{y}|\mathbf{Z},\rho,\tau)$ by the prior density function for ρ and integrate to obtain

$$\frac{\det(\Omega_Z)^{1/2}}{\tau^{|\mathbf{Z}|}\det(\Sigma_Z)^{1/2}}\int \rho^{n/2}e^{-\rho S_Z^2/2}\rho^{\alpha_\rho}e^{-\lambda_\rho\rho}d\rho \propto \frac{\det(\Omega_Z)^{1/2}}{\tau^{|\mathbf{Z}|}\det(\Sigma_Z)^{1/2}}\left(\lambda_\rho + \frac{S_Z^2}{2}\right)^{-\left(\frac{n}{2}+\alpha_\rho\right)},$$

since the integrand is the density function of Gamma $\left(\alpha_\rho + \frac{n}{2}, \lambda_\rho + \frac{1}{2}S_Z^2\right)$, up to a normalizing factor. Hence, the marginal posterior for $\mathbf{Z}$ and τ is given by (3).

Along the lines of Malsiner-Walli and Wagner (2011), we present an approximation of the posterior expected value $\mathbb{E}[Z_v|\mathbf{Z}_{[-v]},\tau,\mathbf{y}]$. If $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are equal except that $Z_v = 0$ and $\tilde{Z}_v = 1$ for one v, then

$$\frac{f_{Z,\tau}(\tilde{\mathbf{Z}},\tau|\mathbf{y})}{f_{Z,\tau}(\mathbf{Z},\tau|\mathbf{y})} = \frac{\Pr(Z_v = 1|\mathbf{Z}_{[-v]},\tau,\mathbf{y})}{\Pr(Z_v = 0|\mathbf{Z}_{[-v]},\tau,\mathbf{y})} = \frac{\mathbb{E}[Z_v|\mathbf{Z}_{[-v]},\tau,\mathbf{y}]}{1 - \mathbb{E}[Z_v|\mathbf{Z}_{[-v]},\tau,\mathbf{y}]}. \tag{7}$$

We will use several assumptions in order to simplify the expression on the left and then solve for $\mathbb{E}[Z_v|\mathbf{Z}_{[-v]},\tau,\mathbf{y}]$. Consider the case when the columns of $\mathbf{X}$ are orthogonal, which implies $\mathbf{X}^T\mathbf{X} \approx nI_p$ and the two choices of Σ_Z are essentially the same. Furthermore, $\langle\mathbf{x}_v,\mathbf{y}\rangle = \langle\mathbf{x}_v,\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\rangle \approx n\beta_v + \langle\mathbf{x}_v,\boldsymbol{\epsilon}\rangle$, which is distributed as $\mathcal{N}(0, \frac{n}{\rho}(Z_v n\tau^2 + 1))$; in this context, distinguishing signal from noise requires $n\tau^2 \gg 1$, so we assume this to be the case. Consequently, $\Omega_Z^{-1} \approx (n + \tau^{-2})I_p \approx nI_p$, which in turn implies S_Z^2 is approximately equal to the residual sum of squares (*RSS*) for the model indicated by $\mathbf{Z}$. Finally, reflecting the results of current GWAS, we assume that the portion of variance explained (PVE) by the loci in consideration is rather small, so *RSS* is not much less than $\mathbf{y}^T\mathbf{y} \approx n$ for any model. If one further chooses $\alpha_\rho = \lambda_\rho \ll \frac{n}{2}$, then

$$\frac{f_{Z,\tau}(\tilde{\mathbf{Z}},\tau|\mathbf{y})}{f_{Z,\tau}(\mathbf{Z},\tau|\mathbf{y})} \approx \frac{1}{\tau}\frac{f_Z(\tilde{\mathbf{Z}})}{f_Z(\mathbf{Z})}\left(\frac{\det(\Omega_{\tilde{Z}})\det(\Sigma_Z)}{\det(\Sigma_{\tilde{Z}})\det(\Omega_Z)}\right)^{1/2}\left(\frac{S_{\tilde{Z}}^2}{S_Z^2}\right)^{-n/2}. \tag{8}$$

Properties of the beta and gamma functions give that the ratio of f_Z values is $(A_\omega + |\mathbf{Z}|)/(B_\omega + p - |\mathbf{Z}| - 1)$. Furthermore, $\sqrt{\det(\Omega_{\tilde{Z}})\det(\Sigma_Z)}/\sqrt{\det(\Sigma_{\tilde{Z}})\det(\Omega_Z)} \approx 1/\sqrt{n}$. Finally, $S^2_{\tilde{Z}} \approx \mathbf{y}^T\mathbf{y} - \frac{1}{n}\sum_{u:Z_u=1}(\mathbf{x}_u^T\mathbf{y})^2$, so

$$\frac{S^2_{\tilde{Z}}}{S^2_Z} \approx \frac{S^2_Z - \frac{1}{n}(\mathbf{x}_v^T\mathbf{y})^2}{S^2_Z} \approx 1 - \frac{(\mathbf{x}_v^T\mathbf{y})^2}{n\mathbf{y}^T\mathbf{y}} \equiv 1 - \eta_v^2.$$

Substituting these results into (7) and (8) gives (4).

B Mathematical details for learning across traits

While the *Unadjusted* prior is not useful, we include its marginal posterior density here for completeness. Its derivation is very similar to that of the basic model, so we focus on the differences. A priori the rows of $\mathbf{Z}$ are independent and each has a beta-binomial distribution, so $f_Z(\mathbf{Z}) = \prod_{v=1}^p \mathrm{B}(A_v + s_v, B_v + q - s_v)/\mathrm{B}(A_v, B_v)$. Furthermore, the columns of $\mathbf{Y}$ are independent given $\mathbf{Z}$, $\boldsymbol{\beta}$, and ρ, and similarly for the columns of $\boldsymbol{\beta}$; so

$$f_{Z,\tau}(\mathbf{Z},\tau|\mathbf{Y}) \propto f_\tau(\tau) f_Z(\mathbf{Z}) \prod_{t=1}^q \left(\lambda_\rho + \frac{S^2_{\mathbf{Z}_t}}{2}\right)^{-\left(\frac{n}{2}+\alpha_\rho\right)} \frac{\det(\Omega_{\mathbf{Z}_t})^{1/2}}{\tau^{|\mathbf{Z}_t|}\det(\Sigma_{\mathbf{Z}_t})^{1/2}}. \tag{9}$$

If $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are equal except that $Z_{vt} = 0$ and $\tilde{Z}_{vt} = 1$ for one v and one t, then the same calculations as for the basic model give that (8) simplifies as

$$\frac{f_{Z,\tau}(\tilde{\mathbf{Z}},\tau|\mathbf{y})}{f_{Z,\tau}(\mathbf{Z},\tau|\mathbf{y})} \approx \frac{1}{\tau\sqrt{n}} \frac{A_v + s_v}{B_v + q - s_v - 1}(1 - \eta_{vt}^2)^{-n/2}.$$

This leads to the approximation (5).

Next we consider the *Across Traits* prior. The posterior density is the same as in (9) except that f_Z is replaced by

$$\begin{aligned} f_{W,Z}(\mathbf{W},\mathbf{Z}) &= \int \Pr(\mathbf{Z}|\mathbf{W},\boldsymbol{\nu})\Pr(\mathbf{W}|\omega_W)\Pr(\boldsymbol{\nu})\Pr(\omega_W)d\boldsymbol{\nu}d\omega_W \\ &= \frac{\mathrm{B}(A_W + |\mathbf{W}|, B_W + p - |\mathbf{W}|)}{\mathrm{B}(A_W, B_W)} \prod_{v:W_v=1} \frac{\mathrm{B}(A_v + s_v, B_v + q - s_v)}{\mathrm{B}(A_v, B_v)}, \end{aligned}$$

provided that $Z_{vt} = 0$ for all v such that $W_v = 0$—otherwise, $f_{W,Z}(\mathbf{W},\mathbf{Z}) = 0$.

To derive the approximation for the posterior expected values, consider $\mathbf{W}$ and $\tilde{\mathbf{W}}$ that are equal except that $W_v = 0$ and $\tilde{W}_v = 1$ for one v. Choose $\mathbf{Z}$ consistent with $\mathbf{W}$, which means $f_{W,Z}(\mathbf{W},\mathbf{Z}) \neq 0$. Since we are using a subscript to denote a column of a matrix, we use a superscript as in $\mathbf{Z}^v$ to denote row v of $\mathbf{Z}$. Furthermore, $\mathbf{Z}^{[-v]}$ denotes the sub-matrix of $\mathbf{Z}$ obtained by deleting row v. Straightforward modification of (7) gives

$$\frac{\sum_{\tilde{\mathbf{Z}}} f_{W,Z,\tau}(\tilde{\mathbf{W}},\tilde{\mathbf{Z}},\tau|\mathbf{Y})}{f_{W,Z,\tau}(\mathbf{W},\mathbf{Z},\tau|\mathbf{Y})} = \frac{\mathbb{E}[W_v|\mathbf{W}_{[-v]},\mathbf{Z}^{[-v]},\tau,\mathbf{Y}]}{1 - \mathbb{E}[W_v|\mathbf{W}_{[-v]},\mathbf{Z}^{[-v]},\tau,\mathbf{Y}]},$$

where the summation is over all $\tilde{\mathbf{Z}}$ such that $\tilde{\mathbf{Z}}^{[-v]} = \mathbf{Z}^{[-v]}$. Furthermore, for any such $\tilde{\mathbf{Z}}$,

$$\frac{f_{W,Z,\tau}(\tilde{\mathbf{W}},\tilde{\mathbf{Z}},\tau|\mathbf{Y})}{f_{W,Z,\tau}(\mathbf{W},\mathbf{Z},\tau|\mathbf{Y})} \approx \frac{A_W + |\mathbf{W}_{[-v]}|}{B_W + p - |\mathbf{W}_{[-v]}| - 1} \frac{\mathrm{B}(A_v + \tilde{s}_v, B_v + q - \tilde{s}_v)}{\mathrm{B}(A_v, B_v)} \prod_{t:\tilde{Z}_{vt}=1} \frac{1}{\tau\sqrt{n}}(1 - \eta_{vt}^2)^{-n/2}.$$

Hence, $\mathbb{E}[W_v|\mathbf{W}_{[-v]}, \mathbf{Z}^{[-v]}, \tau, \mathbf{Y}]^{-1} \approx 1 + 1/h_v(\mathbf{W}_{[-v]}, \tau, \mathbf{Y})$ with

$$h_v(\mathbf{W}_{[-v]}, \tau, \mathbf{Y}) = \frac{A_W + |\mathbf{W}_{[-v]}|}{B_W + p - |\mathbf{W}_{[-v]}| - 1} \sum_{\mathbf{Z}^v} \left[\frac{\mathrm{B}(A_v + |\mathbf{Z}^v|, B_v + q - |\mathbf{Z}^v|)}{\mathrm{B}(A_v, B_v)(\tau\sqrt{n})^{|\mathbf{Z}^v|}} \prod_{t:Z_{vt}=1} (1 - \eta_{vt}^2)^{-n/2} \right],$$

where the summation is over all 2^q possible values of $\mathbf{Z}^v$.

C Mathematical details for learning across sites

The derivation of the marginal posterior for learning *Across Sites* consists of straightforward modifications of previous calculations. The marginal posterior distribution is as in (3) except that the prior $f_Z(\mathbf{Z})$ is replaced by the joint prior for $\mathbf{Z}$ and $\mathbf{G}$, which is

$$f_{G,Z}(\mathbf{G}, \mathbf{Z}) = \frac{\mathrm{B}(A_G + |\mathbf{G}|, B_G + r - |\mathbf{G}|)}{\mathrm{B}(A_G, B_G)} \prod_{\substack{g:G_g=1, \\ p_g>1}} \frac{\mathrm{B}(A_g + s_g, B_g + p_g - s_g)}{\mathrm{B}(A_g, B_g)}$$

provided that $Z_v = 0$ if $G_{\gamma(v)} = 0$, and $Z_v = 1$ if $G_{\gamma(v)} = 1$ and $p_{\gamma(v)} = 1$—otherwise, $f_{G,Z}(\mathbf{G}, \mathbf{Z}) = 0$. Similar to the preceding approximations, $\mathbb{E}[G_g|\mathbf{G}_{[-g]}, \mathbf{Z}_{[-g]}, \tau, \mathbf{y}]^{-1} \approx 1 + 1/h_g(\mathbf{G}_{[-g]}, \tau, \mathbf{y})$, where for $p_g > 1$

$$h_g(\mathbf{G}_{[-g]}, \tau, \mathbf{y}) = \frac{A_G + |\mathbf{G}_{[-g]}|}{B_G + r - |\mathbf{G}_{[-g]}| - 1} \sum_{\mathbf{Z}_g} \left[\frac{\mathrm{B}(A_g + |\mathbf{Z}_g|, B_g + p_g - |\mathbf{Z}_g|)}{\mathrm{B}(A_g, B_g)(\tau\sqrt{n})^{|\mathbf{Z}_g|}} \prod_{\substack{v:\gamma(v)=g, \\ Z_v=1}} (1 - \eta_v^2)^{-n/2} \right]$$

with the summation being over all 2^{p_g} possible values of $\mathbf{Z}_g$. When $p_g = 1$, however, the summation is replaced by $(1 - \eta_v^2)^{-n/2}/(\tau\sqrt{n})$, where v is the lone variant in group g. Hence, the posterior conditional probability of G_g depends upon the overall number of groups and the number of groups considered relevant, while the posterior conditional probability of Z_v given that $G_g = 1$ depends on the number s_g of variants in the same group that are deemed functional.

Supplementary Material

Choice of hyperparameters

We state here the hyperparameters used in most of our computations; the effect of alternate hyperparameters is discussed below. For fine-mapping, there is considerable uncertainty about the likely value of ω, so we use the uninformative prior with $A_\omega = B_\omega = 1$. Since $n \geq 5000$ in our examples, we use $\alpha_\rho = \lambda_\rho = 10$ to ensure that $\alpha_\rho = \lambda_\rho \ll \frac{n}{2}$.

The bounds on the uniform prior for τ justify additional explanation. The lower bound could be zero, but this may lead to increased computational times because very small τ implies small nonzero values in $\boldsymbol{\beta}$ are more likely, which may make the MCMC iteration accept larger models. The assumption $n\tau^2 \gg 1$ suggests choosing $\tau_1 \approx 1/\sqrt{n}$ instead. Furthermore, τ^2/ρ is approximately the variance of the nonzero coefficients since the diagonal entries of either Σ_Z are approximately one. The standardization of $\mathbf{y}$ and the assumption of small PVE implies $\rho \approx 1$. Consequently, τ_2 should be at least three times as large as the largest likely value of $|\beta_v|$. Either marginal or multivariate linear regression of our actual data gives all coefficients with absolute value less than 0.2, so we use $\tau_1 = 0.01$ and $\tau_2 = 10$ in our analyses.

MCMC sampling

We use Gibbs sampling to alternate between updating τ and updating the indicator variables. We use Metropolis-Hastings for each update. We only update the indicator variables once for each update of τ, but it might be more efficient to try multiple indicator variable jumps between each jump of τ.

To define the notation, the Metropolis-Hastings algorithm to generate samples with density proportional to $f(\cdot)$ draws a proposal $\tilde{u}$ from a jumping distribution $Q(\tilde{u}|u^{(m)})$, where $u^{(m)}$ is the previous sample. Next, compute

$$R = \frac{Q\left(u^{(m)}\middle|\tilde{u}\right)}{Q\left(\tilde{u}\middle|u^{(m)}\right)} \frac{f\left(\tilde{u}\right)}{f\left(u^{(m)}\right)}. \tag{10}$$

The new sample is $\tilde{u}$ with probability $\min(R, 1)$ and is $u^{(m)}$ otherwise. To avoid overflow, we actually compute $\log R$ and then take its exponential. The remainder of this section gives the definition of the jumping distributions used.

The jumping distribution for $\tilde{\tau}$ given τ is $\mathcal{N}(\tau, (\tau_2 - \tau_1)^2/16)$ but truncated to the interval (τ_1, τ_2). When $|\mathbf{Z}| = 0$, the marginal posterior density is independent of τ in (τ_1, τ_2). Hence, if all effect sizes are very small, the τ samples may be essentially uniform. This is because the data do not enable learning the effect sizes and has nothing to do with the MCMC algorithm or its convergence.

With one level of indicator variables, our proposal distribution $Q(\tilde{\mathbf{Z}}|\mathbf{Z})$ is nonzero only if $\tilde{\mathbf{Z}} - \mathbf{Z} = \delta \mathbf{e}_v$ for $\delta \in \{-1, +1\}$ and one v, where $\mathbf{e}_v$ is the vector with entry v equal to one and all others equal to zero. Furthermore, for such $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$,

$$Q_Z(\tilde{\mathbf{Z}}|\mathbf{Z}) = \begin{cases} \frac{1}{p} & \text{if } |\mathbf{Z}| = 0 \text{ or } |\mathbf{Z}| = p \\ \frac{\pi_+}{p - |\mathbf{Z}|} & \text{if } \delta = +1 \\ \frac{1-\pi_+}{|\mathbf{Z}|} & \text{if } \delta = -1 \end{cases}, \tag{11}$$

where π_+ is the probability of adding a variable to the model.

Next we define $Q_{W,Z}(\tilde{\mathbf{W}}, \tilde{\mathbf{Z}}|\mathbf{W}, \mathbf{Z})$ for learning *Across Traits*. The first step at each iteration is to choose δ from $\{-1, 0, +1\}$ with probabilities π_-^*, π_0^*, and π_+^*, respectively, with obvious modifications if, say, $\delta = -1$ is not possible because $|\mathbf{W}| = 0$. Then $\tilde{\mathbf{W}} - \mathbf{W} = \delta \mathbf{e}_v$, where

variant v is chosen uniformly from the possible candidates. The proposal $\tilde{\mathbf{Z}}$ will be the same as $\mathbf{Z}$ except for changes to $\mathbf{Z}^v$. If $\delta = -1$, $\tilde{\mathbf{Z}}^v = \mathbf{0}$. If $\delta = 0$, (11) gives $Q_Z^*(\tilde{\mathbf{Z}}^v|\mathbf{Z}^v)$ with q replacing p. If $\delta = +1$, then draw $\tilde{\mathbf{Z}}^v$ based on the prior for $(\mathbf{Z}|\mathbf{W})$: first draw s_v from the beta-binomial distribution

$$\Pr(s_v) = \frac{\mathrm{B}(A_v + s_v, B_v + q - s_v)}{\mathrm{B}(A_v, B_v)} \binom{q}{s_v} \tag{12}$$

and then sample s_v distinct entries uniformly from $\{1, \ldots, q\}$. The overall jumping probability is $Q_{W,Z}(\tilde{\mathbf{W}}, \tilde{\mathbf{Z}}|\mathbf{W}, \mathbf{Z}) = Q_0(\tilde{\mathbf{Z}}|\tilde{\mathbf{W}}, \mathbf{W}, \mathbf{Z})\, Q_W(\tilde{\mathbf{W}}|\mathbf{W})$, where

$$Q_0(\tilde{\mathbf{Z}}|\tilde{\mathbf{W}}, \mathbf{W}, \mathbf{Z}) = \begin{cases} 1 & \text{if } \delta = -1 \\ Q_Z^*(\tilde{\mathbf{Z}}^v|\mathbf{Z}^v) & \text{if } \delta = 0 \\ \frac{\mathrm{B}(A_v+s_v, B_v+q-s_v)}{\mathrm{B}(A_v, B_v)} & \text{if } \delta = +1 \end{cases} \tag{13}$$

and the nonzero values of the mass function for $(\tilde{\mathbf{W}}|\mathbf{W})$ are

$$Q_W(\tilde{\mathbf{W}}|\mathbf{W}) = \begin{cases} \hat{\pi}_W & \text{if } \delta = 0 \\ \frac{1}{p} & \text{if } |\mathbf{W}| = 0 \\ \frac{\pi_-^*}{(1-\pi_+^*)p} & \text{if } |\mathbf{W}| = p \text{ and } \delta = -1 \\ \frac{\pi_+^*}{p-|\mathbf{W}|} & \text{if } \delta = +1 \text{ and } |\mathbf{W}| \neq 0 \\ \frac{\pi_-^*}{|\mathbf{W}|} & \text{if } \delta = -1 \text{ and } |\mathbf{W}| \neq p \end{cases} . \tag{14}$$

The value of $\hat{\pi}_W$ is irrelevant because it will be the same for $Q_W(\tilde{\mathbf{W}}|\mathbf{W})$ and $Q_W(\mathbf{W}|\tilde{\mathbf{W}})$, which will consequently cancel each other in (10). Furthermore, if $\delta = +1$, then $Q_0(\tilde{\mathbf{Z}}|\tilde{\mathbf{W}}, \mathbf{W}, \mathbf{Z})$ in the denominator in (10) will cancel the leftover factor in $f_{W,Z}(\tilde{\mathbf{W}}, \tilde{\mathbf{Z}})$ in the numerator; whereas for $\delta = -1$ a factor in $f_{W,Z}(\mathbf{W}, \mathbf{Z})$ cancels $Q_0(\mathbf{Z}|\mathbf{W}, \tilde{\mathbf{W}}, \tilde{\mathbf{Z}})$.

The mass functions for the *Across Sites* prior are more complicated, but the fortunate cancellations still occur. First, p_g replaces p in (11) to give $Q_Z^*(\tilde{\mathbf{Z}}_g|\mathbf{Z}_g)$ when $\delta = 0$, and (13) becomes

$$Q_0(\tilde{\mathbf{Z}}|\tilde{\mathbf{G}}, \mathbf{G}, \mathbf{Z}) = \begin{cases} 1 & \text{if } \delta = -1 \\ Q_Z^*(\tilde{\mathbf{Z}}_g|\mathbf{Z}_g) & \text{if } \delta = 0 \\ 1 & \text{if } \delta = +1 \text{ and } p_g = 1 \\ \frac{\mathrm{B}(A_g+s_g, B_g+p_g-s_g)}{\mathrm{B}(A_g, B_g)} & \text{if } \delta = +1 \text{ and } p_g > 1 \end{cases} . \tag{15}$$

Let $G_g^* = 1$ if $G_g = 1$ and $p_g > 1$; otherwise it is zero. Assuming that at least one group has $p_g > 1$, the nonzero values of the mass function for $(\tilde{\mathbf{G}}|\mathbf{G})$ are

$$Q_G(\tilde{\mathbf{G}}|\mathbf{G}) = \begin{cases} \hat{\pi}_G & \text{if } \delta = 0 \\ \frac{1}{r} & \text{if } |\mathbf{G}| = 0 \\ \frac{\pi_-^*}{(1-\pi_+^*)r} & \text{if } |\mathbf{G}| = r \text{ and } \delta = -1 \\ \frac{\pi_+^*}{r-|\mathbf{G}|} & \text{if } \delta = +1 \text{ and } |\mathbf{G}^*| \neq 0 \\ \frac{\pi_-^*}{|\mathbf{G}|} & \text{if } \delta = -1, |\mathbf{G}^*| \neq 0, \text{ and } |\mathbf{G}| \neq r \\ \frac{\pi_+^*}{(1-\pi_0^*)(r-|\mathbf{G}|)} & \text{if } \delta = +1, |\mathbf{G}^*| = 0 \text{ and } |\mathbf{G}| \neq 0 \\ \frac{\pi_-^*}{(1-\pi_0^*)|\mathbf{G}|} & \text{if } \delta = -1, |\mathbf{G}^*| = 0, \text{ and } |\mathbf{G}| \neq r \end{cases} . \tag{16}$$

In our experiments, $\pi_+ = 0.5$ for the $\mathbf{Z}$ proposal, while $\pi_0^* = 0.5$ and $\pi_-^* = \pi_+^* = 0.25$ for the $\mathbf{W}$ or $\mathbf{G}$ proposal. One chain starts with no variants in the model while another starts with all variants in the model. For the other two chains, the variants are ranked by the absolute value of their correlation with $\mathbf{y}_t$, and the J_t most strongly correlated have Z_{vt} set to one. For the actual phenotypic data, J_t is 10 for one chain and 20 for the other. For the simulations, one chain has J_t equal to the expected number of variants that are causal for trait t (as determined by the ω or ν used to generate $\mathbf{Z}_t$), and the other chain has J_t double that. The burn-in interval was 10,000 iterations. For the simulated data, 500,000 samples were then used to compute the averages, which took a couple of hours or less (typically much less, especially when using only one trait at a time) running the chains sequentially. For the actual data, 10 times as many samples were used for the averages; the chains were run in parallel and each MCMC sampling took 2–3 hr. We always used the g-prior.

To assess convergence of the averages $\overline{Z}_{vt}$, we computed the average for each chain and then set $\Delta\overline{Z}_{vt}$ equal to the difference between the maximum and minimum averages over the four chains. For each prior applied to each set of 100 datasets, at least 95% of $\Delta\overline{Z}_{vt}$ were less than 0.05. For the priors that incorporate $\mathbf{W}$ or $\mathbf{G}$, at least 95% of the analogous values of $\Delta\overline{W}_v$ or $\Delta\overline{G}_g$ were less than 0.05 except that only 92% of $\Delta\overline{G}_g$ satisfied this condition when the *Across Sites* prior was applied to the traits simulated from the genotype data.

Evaluation of variable selection performance

The p-values for use in the BH procedure are from ordinary least squares, using either one variant at a time ('marginal') or all variants simultaneously ('full'). For each dataset, we apply BH to all traits simultaneously, which means each BH test has pq hypotheses. For the performance curves, we choose a target FDR, say α, and apply BH with that α to each dataset, computing the false discovery proportion (FDP) and of discoveries of true null hypotheses for each dataset. We then average over the 100 datasets to obtain the FDR and power for one point in our plots. For our performance curves, we used the FDR control targets $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.3, 0.5\}$. Applying BH to the traits separately gave essentially the same results for the simulated data; the effect of testing traits separately in the actual data is discussed below.

The performance curves for the Lasso show the effect of varying the penalty parameter λ. We use the R function `glmnet` with its default parameters (with `family='gaussian'`) except without an intercept. This returns the fits for a set of λ values, but that set depends upon the input data. We compute the FDP and power at each value of λ output for that trait. Using `locpoly()` in the R package `KernSmooth`, we perform local linear regression on all these points for all datasets to approximate FDP and power as functions of λ. When instead choosing a specific λ by cross-validation, the one-standard error rule gives the largest value of λ such that the cross-validation error is within one standard error of the minimum error. In this case, the choice depends upon the randomly drawn cross-validation bins, so for reproducibility we set the random seed to 1234 immediately before calling `cv.glmnet`.

To compute the performance curves for Bayesian selection of variants, we choose threshold ξ, select using the rule $\overline{Z}_{vt} > \xi$ for each dataset, compute the FDP and power for each dataset (same as for BH), and then average over all 100 datasets to obtain one point in our plots. The values of ξ were 0.01ℓ, $1 - 0.01\ell$, and 0.1ℓ for $\ell = 1, \ldots, 9$.

Genotype and phenotype data

Variants in dbSNP version 137 are indicated with their *rs* names, while other variants are referred to as *v_c9_107548661*, say, specifying chromosome and position (from GRCh37). Starting with the genotype and phenotype data in Service *et al.* (2014), we removed 786 individuals that were missing values for any of the principal components or the three traits HDL, LDL, and

log-transformed TG, leaving 5335 subjects. From the sequencing data that had passed quality control, we removed two variants that were missing data for over 20% of the subjects as well as variants whose minor allele occurred only once in the remaining subjects, leaving 1326 variants.

Additional filtering involved: (1) annotation, whose possible values were nonsense, missense, coding synonymous, UTR 5′, UTR 3′, intron, and unknown; (2) PolyPhen2 predictions, whose possible values were probably damaging, possibly damaging, benign, and unknown; (3) marginal regression p-values from Service *et al.* (2014); and (4) minor allele frequency (MAF). If all subjects have exactly the same genotype (including missing values) for two or more variants, we chose one of the variants using the ad hoc rules in Figure 7, thereby eliminating 24 more variants. This $\mathbf{X}$ was used in the exact computation of the posterior probabilities in individual loci discussed briefly at the end of the Results section of the main paper.

For the MCMC method of estimating the posterior probabilities across all loci simultaneously, we further pruned to obtain a set of variants with maximal correlation less than a specified bound $C_{\max}$, for reasons discussed in the main paper. Figure 8 gives our algorithm to achieve this. The set U_0 contains variants mentioned in Service *et al.* (2014) and Bogdan *et al.* (in press). It was introduced so that we only dropped such variants when they were correlated with each other, which simplified comparison with those papers. Furthermore, we manually chose which of two variants to drop in some cases, as we now explain. Service *et al.* (2014) list 39 associations between an 'Array SNP' and a trait. Of those, 16 had p-value less than 0.001 for their data and involved one of the three traits HDL, LDL or TG. Our filtering process dropped one of these: *rs10096633*, which was an Array SNP for both HDL and TG, because it had correlation 0.96 with *rs328*, which has a nonsense mutation. Among the new discoveries reported by Service *et al.* (2014), *rs651821* and *rs2266788* in gene APOA5 are both associated with one of our traits, but we had to drop one of them because their correlation is 0.95. We dropped the latter because it had a slightly stronger correlation with the Array SNP *rs12805061*.

For almost all of our results, $C_{\max} = 0.3$. In this case, after filtering, we put *rs12805061* and *rs3135506* (which is a missense mutation predicted to be probably damaging) back into $\mathbf{X}$, which then had 768 variants. After imputation, all pairwise correlations in $\mathbf{X}$ are less than 0.3 except for two that are slightly greater. There are 628 variants with MAF<0.01. To explore the effects of pruning the variants, we also consider $\mathbf{X}$ obtained by setting $C_{\max}$ to 0.5, 0.7 and 0.9, resulting in 968, 1042 and 1124 variants, respectively.

Using the phenotype data (via the p-values) when filtering the variants could affect the accuracy of methods to estimate or control FDR, but our previous experience has been that this does not occur in this situation.

Because substituting means for missing genotype values is not accepted practice, we comment on our use of this approach. Since we discard the two variants with very many missing values, only 0.04% of the values in $\mathbf{X}$ (with $C_{\max} = 0.3$) are missing, and only six variants are missing data for more than 1% of subjects—one variant is missing data for 2.2% of subjects and the next worst is missing 1.3%. With $C_{\max} = 0.9$, only 0.08% of the values in $\mathbf{X}$ are missing.

Each of the traits (or its logarithm) were regressed on age, age^2, and indicator variables for sex, oral contraceptive use, pregnancy status, and cohort. The five genetic principal components along with the intercept were regressed out of both $\mathbf{X}$ and the residuals $\mathbf{Y}$ from this regression, and the columns of both were then standardized.

Simulation scenarios

For the datasets with orthogonal $\mathbf{X}$, we set $n = 5000$, $p = 50$, $q = 5$, and $\mathbf{X} = \sqrt{\frac{n-1}{n/p}}(I_p\ I_p\ \cdots\ I_p)^T$ so that $\mathbf{X}^T\mathbf{X} = (n-1)I_p$. The columns of $\mathbf{X}$ were not centered because that would destroy the orthogonality. In generating phenotypic data, $\rho = 1$, $\tau \sim \text{Unif}(0.045, 0.063)$, nonzero $\beta_{vt} \sim$

For each of the 23 sets of variants for which all subjects have the same genotype
    If they are on the same chromosome and have the same annotation and PolyPhen2 prediction
        Choose one arbitrarily
    Else if exactly one variant has annotation 'missense' (none had 'nonsense')
        Choose it
    Else if no variant has annotation 'missense'
        Choose the one with annotation 'coding synonymous' (which meant discarding UTR 3′ variant or variant without annotation)
    Else (this case only occurred once)
        Choose the one with PolyPhen2 prediction of 'probably damaging' (which meant discarding the variant predicted to be benign)

Figure 7: Pseudocode for eliminating duplicate variants.

For mode = 1,2
    For $c = 0.9, 0.8, \ldots, C_{\max}$
        variants = var.filter(variants, mode, c)

Function var.filter(V, m, c_0)
C is the absolute value of the correlation matrix of V, computed with the R option `use="pairwise.complete.obs"`
Consider the entries of C_{ij} in decreasing order
    If $C_{ij} \leq c_0$, return V
    $V^* = \{v_i, v_j\}$
    If (m == 1)
        If v_i and v_j are on different chromosomes, continue
        Add to V^* all v_ℓ on same chromosome as v_i and such that $C_{i\ell}$ or $C_{j\ell}$ is greater than c_0
    U_0 contains any entries in V^* that we particularly want to keep (see text)
    $U_1 = \arg\min_{v_i \in V^*}$ p-value—but only consider common variants
    U_2 contains variants in V^* with the 'worst' annotation, where the ordering is nonsense, then missense with probably damaging, then missense with possibly damaging, then missense with any prediction
    If U_0 is not empty
        U is the entry in U_0 with the highest priority
    Else if U_1 is empty
        $U = U_2$
    Else if U_2 is empty
        $U = U_1$
    Else
        $U = U_1 \cap U_2$
        If U is empty
            $U = U_1$
    If U is empty, $U = V^*$
    Stop in the following rules when one gives single variant u:
        1. $u \in U$
        2. u is variant in U with greatest MAF
        3. Choose u from U arbitrarily
    Drop u from V and C

Figure 8: Pseudocode for filtering correlated variants.

$\mathcal{N}(0, \tau^2)$, and the distributions of the probabilities of association were as follows:

Exchangeable variants one $\omega \sim \text{Beta}(12, 48)$ was drawn independently for each trait.

Pleiotropy one $\omega_W \sim \text{Beta}(16, 55)$ was drawn for each dataset, W_v was drawn for each variant, and then $\nu_v \sim \text{Beta}(48, 12)$ was drawn for each nonzero W_v.

Gene effect $\omega_G \sim \text{Beta}(16, 55)$ was drawn independently for each trait, G_g was drawn for each group of five consecutive variants, and then $\nu_g \sim \text{Beta}(48, 12)$ was drawn for each nonzero G_g.

These choices ensured $n\tau^2 \gg 1$ so that signal was greater than noise, $\mathbb{E}[Z_{vt}] \approx 0.2$, and $|\mathbf{Z}_t|$ exceeded 15 only about 5–15% of the time. Very roughly

$$\langle \mathbf{y}_t, \mathbf{y}_t \rangle = \langle \mathbf{X}_{\mathbf{Z}_t} \boldsymbol{\beta}_{\mathbf{Z}_t}, \mathbf{X}_{\mathbf{Z}_t} \boldsymbol{\beta}_{\mathbf{Z}_t} \rangle + 2\langle \mathbf{X}_{\mathbf{Z}_t} \boldsymbol{\beta}_{\mathbf{Z}_t}, \boldsymbol{\epsilon}_t \rangle + \langle \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t \rangle \approx n\langle \boldsymbol{\beta}_{\mathbf{Z}_t}, \boldsymbol{\beta}_{\mathbf{Z}_t} \rangle + \langle \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t \rangle \approx \frac{n}{\rho} \left(\tau^2 |\mathbf{Z}_t| + 1 \right),$$

so the generated traits also had variance close to one—between 0.94 and 1.17. There were 100 datasets, each with $q = 5$ traits. When applying the *Across Sites* prior to any of these data sets, the groups are the same as when generating the dataset with gene effect: the first group consists of the first five variants, the second group consists of the next five variants, and so forth.

When simulating phenotypes from the actual genotype data, all rare (MAF<0.01) variants in the same gene with missense or nonsense mutations are in the same group. Every other variant—common or rare—is in a group by itself. Then $\mathbf{Z}$ was generated randomly as follows:

- for each trait t, draw one group from amongst those with at least five variants and set $Z_{vt} = 1$ for 3–4 rare variants (number drawn uniformly) in that group
- draw 10 common variants and, for every trait t, set $Z_{vt} = 1$ with probability 0.9
- draw 40 common variants and, for every trait t, set $Z_{vt} = 1$ with probability 0.1

The rest of the process to generate each trait was the same as in the other simulations. Again, there were 100 datasets, each with $q = 3$ traits. There are 20 genes with at least five nonsynonymous rare variants; 16 of them have fewer than 10 variants, but the others have 14–16 variants.

We also created another set of simulations along the lines of a traditional investigation of pleiotropy versus coincident linkage, which we will call the pleiotropy-linkage experiment. We began with $\mathbf{X}$ created by setting $C_{\max} = 0.7$ in Figure 8. We then took pairwise correlations between all variants with MAF >0.1. From the 10 pairs with absolute correlation greater than 0.6 and in the same locus, we selected the five pairs listed in Table 3. For each selected pair, we simulated 40 datasets, each with $\mathbf{X}$ restricted to the one locus and with two traits. In half the datasets, one of the variants is causal for both traits (pleiotropy, with each variant causal equally often) while the other datasets have one variant causal for one phenotype and the other variant causal for the other phenotype (coincident linkage). No other variants were causal. The rest of the process to generate each trait was the same as in the other simulations.

Results for illustrative example

For our Bayesian variable selection methods and several non-Bayesian methods applied to the datasets with orthogonal $\mathbf{X}$, Figure 9 compares the power for FDR less than 0.2. In this range, the power of the *Across Sites* prior applied to the dataset with pleiotropy is only slightly less than that of the basic prior or the non-Bayesian methods. Similarly, the loss of power of the *Across Traits* prior applied to the dataset exhibiting gene effects is slight when FDR is less than 0.2.

Table 3: Pairs of variants used in the pleiotropy-linkage experiment. The total number of variants in the datasets for that pair is p.

variants (MAF)				locus	p	correlation
rs11142	(0.29)	rs464218	(0.45)	CELSR2	128	0.70
rs567243	(0.49)	rs560887	(0.31)	G6PC2	62	-0.66
rs15285	(0.24)	rs316	(0.12)	LPL	25	0.63
rs12445698	(0.19)	rs5801	(0.18)	CETP	88	0.65
rs1801706	(0.21)	rs5882	(0.39)	CETP	88	0.66

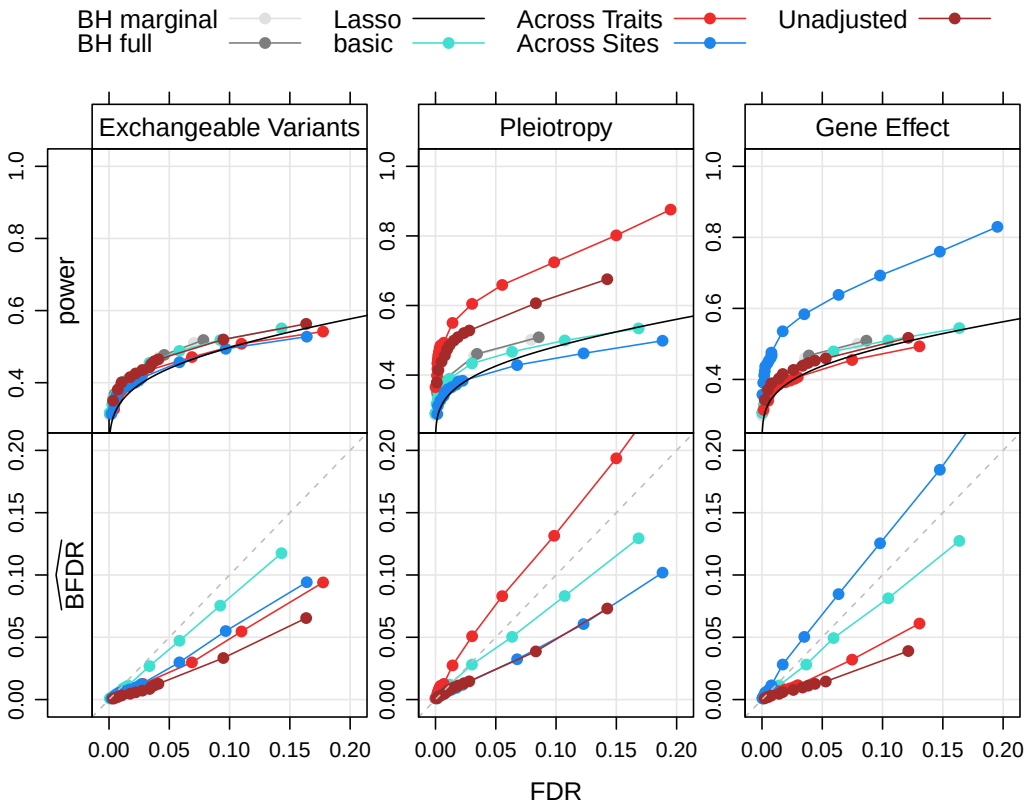

Figure 9: Power (top) and $\widehat{\text{BFDR}}$ (bottom) as a function of empirical FDR in the illustrative example, limiting to FDR<0.2. Each color indicates a different variable selection approach (see legend at the top). The different columns correspond to different data generating mechanisms.

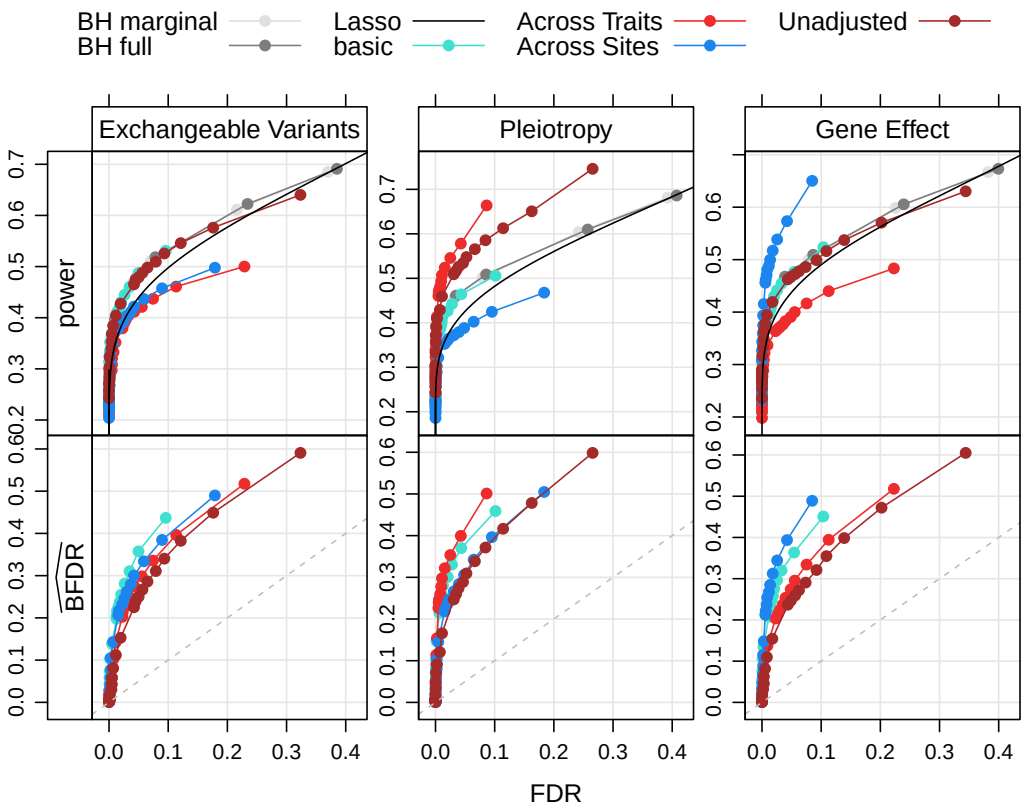

Figure 10: Power (top) and $\widehat{\text{BFDR}}$ (bottom) as a function of empirical FDR in the illustrative example with prior $\tau \sim \text{Unif}(2, 5)$. Each color indicates a different variable selection approach (see legend at the top). The different columns correspond to different data generating mechanisms.

We also investigated the effect of using other hyperparameters. Figure 10 shows the results when the prior for τ was $\text{Unif}(2, 5)$ but everything else was the same as for the results in Figure 9. As suggested by the approximations given in the main paper for the conditional posterior expected values of the indicator variables, increasing τ causes $\overline{Z}_{vt}$ to decrease; so fewer variables are selected for given ξ, decreasing both FDR and power. This also means that $\overline{Z}_{vt}$ underestimates the probability that variant v is causal for trait t, and $\widehat{\text{BFDR}}$ greatly overestimates FDR. On the other hand, using the prior $\tau \sim \text{Unif}(0.045, 0.063)$ as was used in data generation gave essentially the same results as in Figure 9 because the mean and variance of the τ samples were usually less than 0.1 when using the uninformative prior $\tau \sim \text{Unif}(0.01, 10)$.

Finally, we tried using different hyperparameters for the priors on the probabilities of causality but with everything else the same as for the results in Figure 9. For the dataset with exchangeable variants, we again used uninformative priors for ω_W and ω_G but all other Beta priors had parameters $A = 12$ and $B = 48$. For the dataset with pleiotropy, we changed only the hyperparameters for ω_W and $\boldsymbol{\nu}$ for the *Across Traits* prior, which now have the same values as during data generation, and likewise for the dataset with gene effects and the *Across Sites* prior. Figure 11 shows that this did not have much effect on the results. This suggests there is not much advantage to using tighter priors, even if it were possible to guess the hyperparameters so accurately.

Results for generating phenotypes from actual genotype data

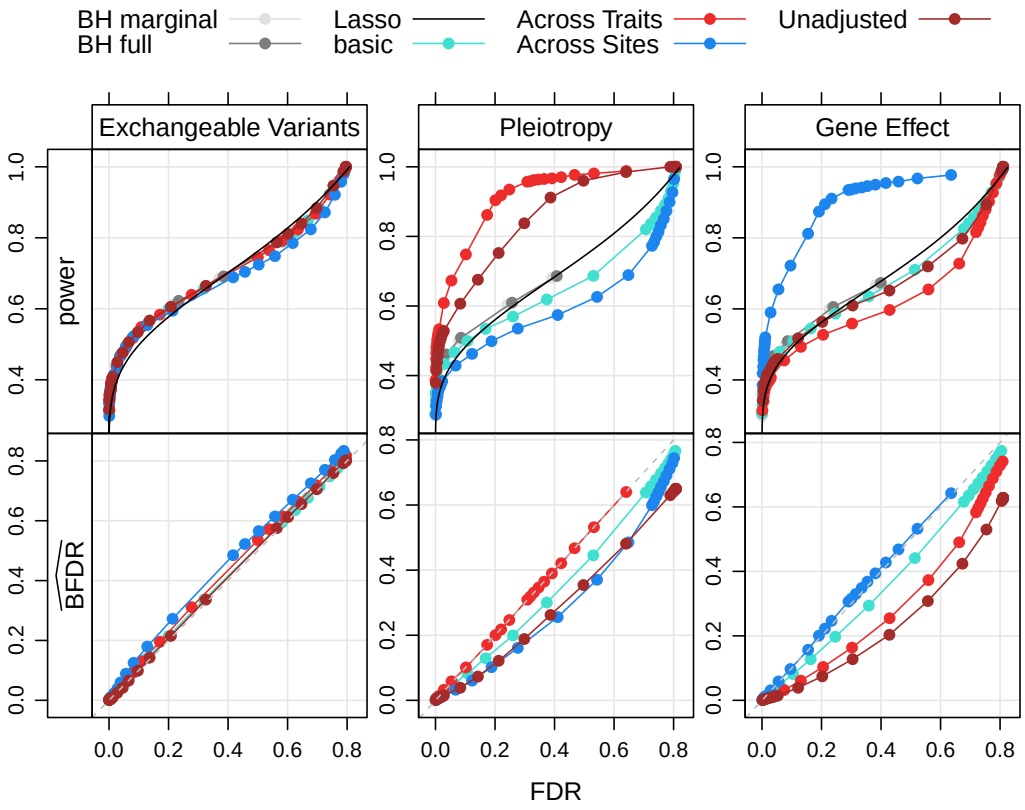

Figure 11: Power (top) and $\widehat{\text{BFDR}}$ (bottom) as a function of empirical FDR in the illustrative example with tighter priors for the probability of causality in some cases. Each color indicates a different variable selection approach (see legend at the top). The different columns correspond to different data generating mechanisms.

Table 4: FDR and power for selected values of ξ when applying Bayesian methods to simulated phenotypes with actual genotype data.

	FDR	power	FDR	power	FDR	power
	$\xi = 0.5$		$\xi = 0.7$		$\xi = 0.8$	
basic	0.100	0.41	0.039	0.37	0.021	0.35
Across Traits	0.133	0.51	0.057	0.45	0.031	0.42
Across Sites	0.143	0.46	0.054	0.41	0.029	0.38

When applying the *Across Sites* prior to data with actual genotype data, the groups were the same as described above for simulating phenotypes from this data. This mimicked the burden tests in Service *et al.* (2014).

To create the performance curves in the lower row of plots in Figure 5 in the main paper, the FDR for each ξ is equal to that computed for the curves for all variants, but the power is computed separately for each type of variant. Table 4 shows the FDR and power when using specified values of ξ to select variants after applying the Bayesian priors to the same set of simulations. The results when $\xi = 0.7$ are essentially the same as the results in the paper from controlling $\widehat{\text{BFDR}} \leq 0.05$.

Next we discuss the results of the pleiotropy-linkage experiment. Over all the datasets, at least 95% of $\Delta\overline{Z}_{vt}$ were less than 0.05 after 500,000 MCMC iterations, and when using the *Across Traits* prior at least 90% of $\Delta\overline{W}_v$ were. Figure 12 shows the FDR and power for the basic and *Across Traits* priors as well as BH applied to both sets of p-values. In the case of separate causal variants, the *Across Traits* prior may have a slight loss of power when FDR is less than 0.5 but is still much better than BH with p-values from the full model. In the case of pleiotropic traits, however, the *Across Traits* prior clearly has greater power per FDR than the other methods, which have roughly the same power regardless of whether the traits have the same or different causal variants. For the variants other than the pair listed in Table 3, 8% of the time the difference in $\overline{Z}_{vt}$ between the two priors is less than 0.05. When the traits are pleiotropic, the differences are similar for the non-causal variant; but for the causal variant, $\overline{Z}_{vt}$ for the non-causal variant is sometimes greater than 0.7 with the *Across Traits* prior even when it is less than 0.2 with the basic prior, consequently increasing power considerably. When the traits are not pleiotropic, *Across Traits* may inflate $\overline{Z}_{vt}$ for the wrong variant of the pair, thereby increasing FDR slightly for given ξ, but very rarely increases $\overline{Z}_{vt}$ by much for the causal variant; so power is not increased.

Results for the case study

Figures 13 and 14 illustrate the model selection results for LDL and TG, respectively, which are interpreted in the same way as the figure for HDL in the main paper. Tables 5–7 list the methods that select each variant for association with HDL, LDL, and TG, respectively, for variants selected by at least one method. An x in a column indicates that the method selects that variant for that trait. The column labeled 'position' gives the position (from GRCh37) of the variant on its chromosome.

For LDL and TG, no method identifies any rare variant, and the *Across Sites* selections always agree with the majority of approaches. The *Across Traits* approach selects a few variants that are noteworthy. For LDL, it selects a variant in each of three loci—CETP, GALNT2, and NCAN—where no other method identifies any signal. In each case, the variant that the *Across Traits* approach selects has a strong association with either HDL or TG. Two of these three loci have previous evidence of association with LDL. For TG, the *Across Traits* approach selects a variant in FADS1, where no other method selects any. This is the same variant that caught our attention

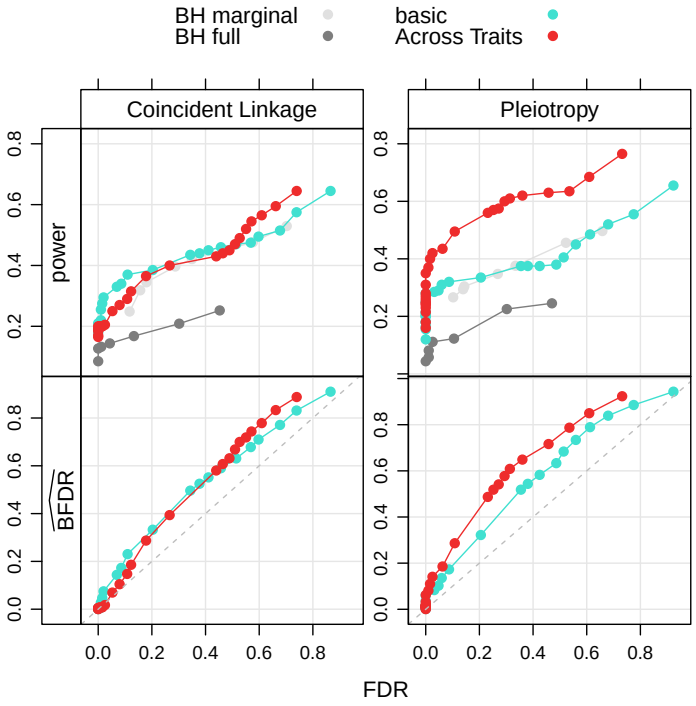

Figure 12: Power (top) and $\widehat{\text{BFDR}}$ (bottom) as a function of empirical FDR in the experiment to check whether *Across Traits* can distinguish between pleiotropy (right column) and traits with different correlated causal variants (left column). Each color indicates a different variable selection approach (see legend at the top).

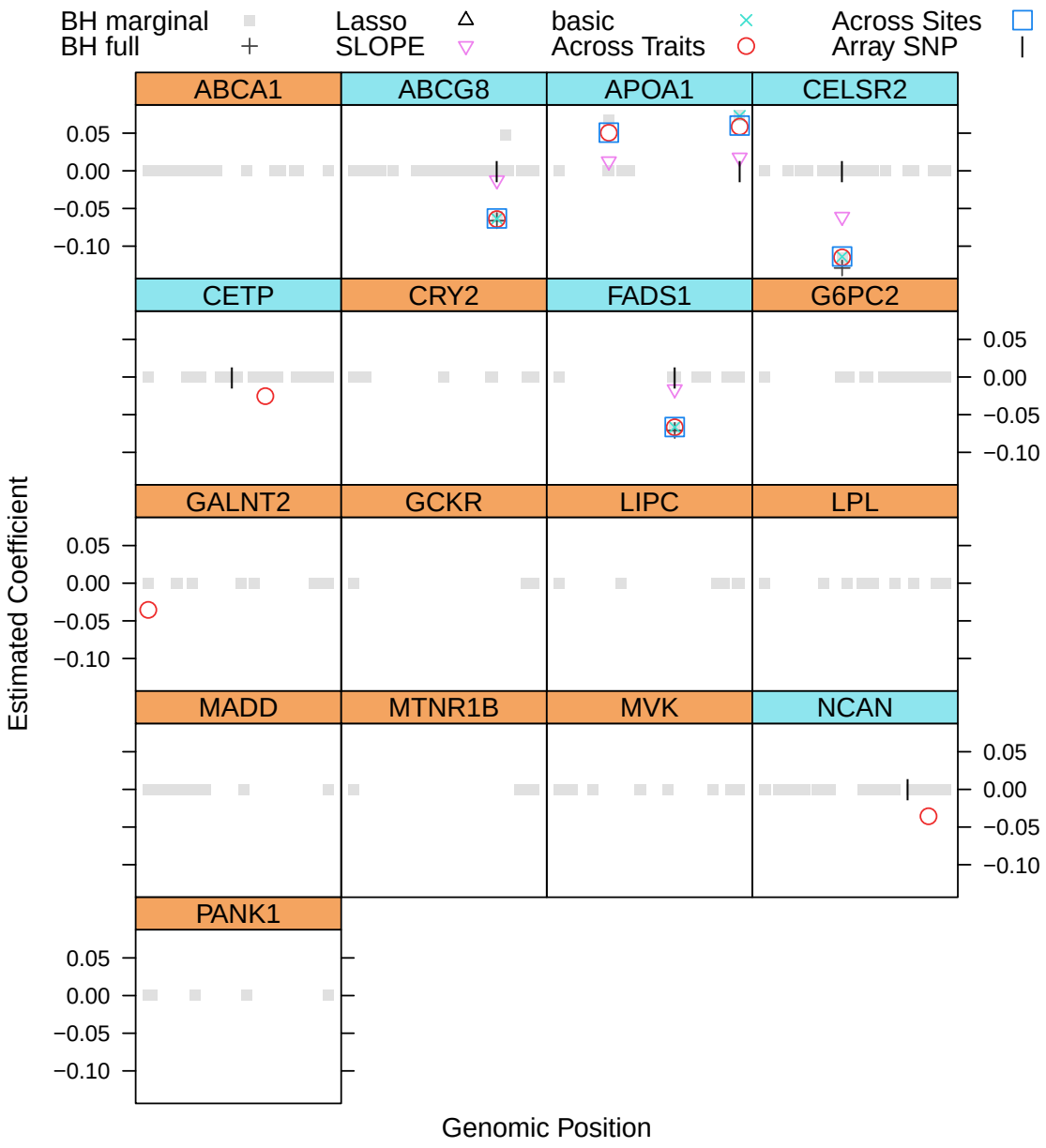

Figure 13: Estimated variant effects on LDL. Each panel corresponds to a locus, the x-axis indicate the variant's genomic position and the y-axis its regression coefficient (with the exception of BH marginal, only nonzero coefficients are represented). The color code of the panel titles indicates the presence/absence of prior evidence of association between the locus and LDL (turquoise/orange, respectively). Model selection methods are distinguished using plotting symbols, as indicated in the legend at the top.

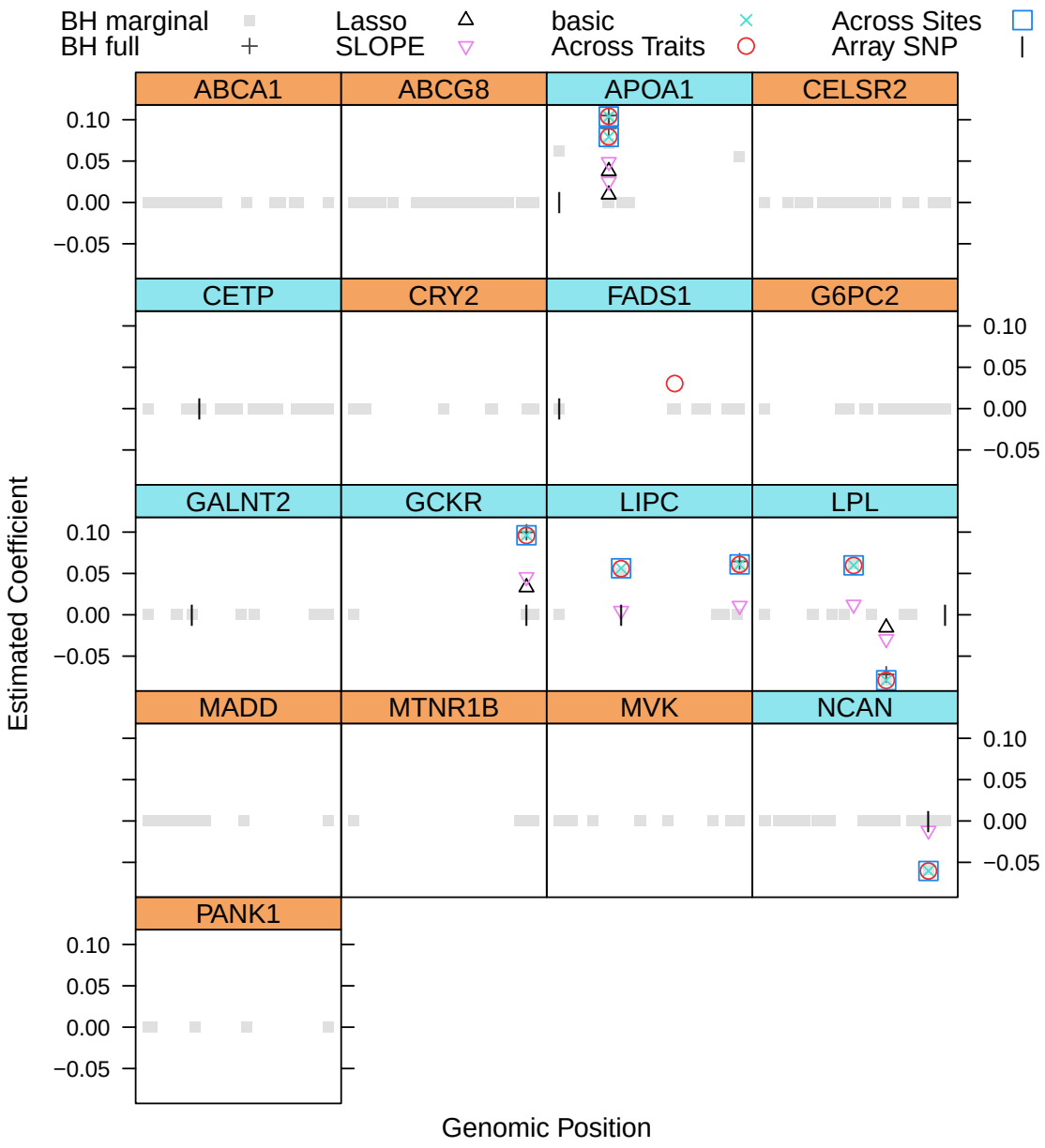

Figure 14: Estimated variant effects on TG. Each panel corresponds to a locus, the x-axis indicate the variant's genomic position and the y-axis its regression coefficient (with the exception of BH marginal, only nonzero coefficients are represented). The color code of the panel titles indicates the presence/absence of prior evidence of association between the locus and TG (turquoise/orange, respectively). Model selection methods are distinguished using plotting symbols, as indicated in the legend at the top.

Table 5: Comparison of selection results for HDL.

variant	position	locus	MAF	BH full	BH marg	Lasso	SLOPE	Across basic	Across Trait	Across Sites	PLOS
v_c9_107548661	107548661	ABCA1	0.0002	x	x	x	x	x		x	x
v_c9_107555091	107555091	ABCA1	0.0005		x		x	x		x	x
v_c9_107555452	107555452	ABCA1	0.0002	x	x	x	x	x	x	x	x
rs76881554	107578620	ABCA1	0.0043							x	
rs2066715	107588033	ABCA1	0.0551		x	x	x	x	x	x	x
rs2066718	107589255	ABCA1	0.0148	x	x		x	x	x	x	x
rs145183203	107646756	ABCA1	0.0029							x	
rs2575875	107662494	ABCA1	0.2999		x	x	x	x	x	x	x
v_c9_107665945	107665945	ABCA1	0.0002							x	
rs12805061	116553025	APOA1	0.2718		x		x				x
rs651821	116662579	APOA1	0.0845		x	x	x	x	x	x	
rs646776	109818530	CELSR2	0.2171						x		
rs34216426	56913088	CETP	0.0002		x		x	x		x	
rs5801	56913513	CETP	0.1818		x		x	x	x	x	
rs5802	56919235	CETP	0.1194		x						
rs3764261	56993324	CETP	0.2719	x	x	x	x	x	x	x	x
rs5883	57007353	CETP	0.0401	x		x	x	x	x	x	x
rs5880	57015091	CETP	0.0242	x	x	x	x	x	x	x	x
rs2303790	57017292	CETP	0.0006	x			x	x	x	x	x
v_c16_57095439	57095439	CETP	0.0002		x		x				
rs509360	61548559	FADS1	0.3364								x
rs174546	61569830	FADS1	0.4212						x		
rs611229	230324067	GALNT2	0.4077		x	x	x	x	x	x	x
rs1532085	58683366	LIPC	0.4431	x	x	x	x	x	x	x	x
rs261336	58742418	LIPC	0.2302	x	x	x	x	x	x	x	
rs28933094	58855748	LIPC	0.0150	x	x	x	x	x	x	x	x
rs268	19813529	LPL	0.0180	x	x	x	x	x	x	x	x
rs328	19819724	LPL	0.0882	x	x	x	x	x	x	x	x
rs10096633	19830921	LPL	0.0943								x
v_c11_47400044	47400044	MADD	0.0005	x	x	x	x	x		x	
rs7946766	48004369	MADD	0.1965		x	x	x	x	x	x	x
rs12314392	110010866	MVK	0.4033							x	x

Table 6: Comparison of selection results for LDL.

variant	position	locus	MAF	BH full	BH marg	Lasso	SLOPE	Across basic	Across Trait	Across Sites	PLOS
rs6756629	44065090	ABCG8	0.0877	x	x		x	x	x	x	x
rs145756111	44102302	ABCG8	0.0115		x						x
rs651821	116662579	APOA1	0.0845		x		x		x	x	x
rs11216267	116952392	APOA1	0.4578		x		x	x	x	x	x
rs646776	109818530	CELSR2	0.2171	x	x		x	x	x	x	x
rs12445698	56928216	CETP	0.1887								x
rs3764261	56993324	CETP	0.2719						x		
rs174546	61569830	FADS1	0.4212	x	x		x	x	x	x	x
rs611229	230324067	GALNT2	0.4077						x		
rs12610185	19721722	NCAN	0.0617								x
rs2304130	19789528	NCAN	0.0610						x		

Table 7: Comparison of selection results for TG.

variant	position	locus	MAF	BH full	BH marg	Lasso	SLOPE	Across basic	Across Traits	Across Sites	PLOS
rs12805061	116553025	APOA1	0.2718		x						x
rs2266788	116660686	APOA1	0.0891								x
rs3135506	116662407	APOA1	0.0597	x	x	x	x	x	x	x	x
rs651821	116662579	APOA1	0.0845	x	x	x	x	x	x	x	
rs11216267	116952392	APOA1	0.4578		x						
rs1561140	56864398	CETP	0.4762								x
rs6591657	61434532	FADS1	0.1506								x
rs174546	61569830	FADS1	0.4212						x		
rs4846930	230346829	GALNT2	0.4050								x
rs1260326	27730940	GCKR	0.3566	x	x	x	x	x	x	x	x
rs261336	58742418	LIPC	0.2302		x		x	x	x	x	x
rs28933094	58855748	LIPC	0.0150	x	x		x	x	x	x	x
rs268	19813529	LPL	0.0180		x		x	x	x	x	x
rs328	19819724	LPL	0.0882	x	x	x	x	x	x	x	x
rs10096633	19830921	LPL	0.0943								x
rs2304130	19789528	NCAN	0.0610		x		x	x	x	x	x

when only the *Across Traits* approach selected it for HDL. Again, this locus has previous evidence of association with TG.

To conclude, we note that applying BH to each phenotype separately gives different results in some cases. When applied to the p-values from the full model for HDL only, BH selects four variants that it does not select when applied to all traits simultaneously: *rs2066715* in ABCA1, *rs651821* in APOA1, *rs34216426* in CETP, and *rs611229* in GALNT2. All of these are also selected for HDL by most other methods. When applied to marginal p-values for HDL only, BH selects three variants that it does not select when applied to all traits simultaneously: *rs62136410* in ABCG8, *rs2303790* in CETP, and *rs11988* in MADD. Most methods select *rs2303790* for HDL, but no other methods select the other two of these; and ABCG8 does not even have prior evidence of association to HDL. The final difference arising from applying BH to traits separately is that it does not select *rs145756111* in ABCG8 for association with LDL when using marginal p-values, which brings it into agreement with the other methods on this variant.

Effect of pruning variants

For the case study, we pruned the set of variants so that the maximum of the absolute values of the pairwise correlations is 0.3 in order to facilitate comparison of the selection results by different methods. We will now describe some experiments we performed with the actual data to investigate the effects of this pruning. First, we compare our selection results presented in the main text and above to the results when using $\mathbf{X}$ obtained by setting $C_{\max}$ in Figure 8 equal to 0.5, 0.7 and 0.9. For clarity, let $\mathbf{X}^{(3)}$, $\mathbf{X}^{(5)}$, $\mathbf{X}^{(7)}$ and $\mathbf{X}^{(9)}$ denote the different versions of $\mathbf{X}$. After 5,000,000 MCMC iterations for each of four chains, over 99% of $\Delta\overline{Z}_{vt}$ and of $\Delta\overline{W}_v$ or $\Delta\overline{G}_g$ are less than 0.05 for all levels of pruning, so convergence still seems acceptable.

The values of $\overline{Z}_{vt}$ usually do not change very much for different versions of $\mathbf{X}$; and when they do, they almost always decrease. More precisely, much less than 1% of the values of $\overline{Z}_{vt}$ *increase* by more than 0.05. For given prior and trait, the percentage of values that *change* by more than 0.05 is less than 3% between $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(5)}$ and 1% between $\mathbf{X}^{(5)}$ and $\mathbf{X}^{(7)}$ and between $\mathbf{X}^{(7)}$ and $\mathbf{X}^{(9)}$.

Table 8 compares the total number of variants selected by our Bayesian priors for each of the

Table 8: Comparison of selection results from BH and Bayesian methods applied to different levels of variant pruning. The columns labeled R give the number of variants selected, and the columns labeled $\widehat{\text{BFDR}}$ report the Bayesian FDR. The selection threshold for the Bayesian methods was $\xi = 0.7$. For BH, $\alpha = 0.05$.

	HDL		LDL		TG	
prior	R	$\widehat{\text{BFDR}}$	R	$\widehat{\text{BFDR}}$	R	$\widehat{\text{BFDR}}$
maximum correlation 0.3						
basic	21	0.046	4	0.012	8	0.011
Across Traits	19	0.084	8	0.059	9	0.029
Across Sites	25	0.064	5	0.063	8	0.007
BH full	13		3		5	
BH marginal	22		6		10	
maximum correlation 0.5						
basic	21	0.068	4	0.014	9	0.024
Across Traits	18	0.091	8	0.060	10	0.044
Across Sites	25	0.076	4	0.012	9	0.020
maximum correlation 0.7						
basic	20	0.067	4	0.014	9	0.033
Across Traits	17	0.081	8	0.057	9	0.022
Across Sites	24	0.071	4	0.013	9	0.029
maximum correlation 0.9						
basic	19	0.073	4	0.038	9	0.050
Across Traits	15	0.071	8	0.090	9	0.048
Across Sites	22	0.071	4	0.037	9	0.045
BH full	6		2		4	
BH marginal	32		13		15	

pruning levels; they still did not select any variants that are in a locus (CRY2, G6PC2, MTNR1B, and PANK1) lacking any prior evidence of association to lipid traits. The table also shows the number of variants selected by applying BH to the p-values from $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(9)}$. For a more detailed look at the differences, Table 9 lists all cases in which a variant would be selected for one version of $\mathbf{X}$ but not another. These differences are explained in detail at the end of this section, but for now we categorize them as follows: (1) one variant (*rs2066715*) is replaced in the selections by another strongly correlated variant (*rs2853579*) when it enters $\mathbf{X}^{(7)}$; (2) two rare variants are not in $\mathbf{X}^{(3)}$ but are selected (or close to the threshold) for the other versions of $\mathbf{X}$, at least in part because one of the 2–3 subjects with the minor allele has an extreme value for the trait; (3) two variants whose $\overline{Z}_{vt}$ is barely above the selection threshold for some trait and prior with $\mathbf{X}^{(3)}$ but then decreases by as much as 0.15 as correlated variants enter $\mathbf{X}$ with $\overline{Z}_{vt} < 0.1$; and (4) three variants that are selected for $\mathbf{X}^{(3)}$ but their $\overline{Z}_{vt}$ decreases by as much as 0.5 as correlated variants enter $\mathbf{X}$ with $\overline{Z}_{vt} > 0.15$.

To summarize, the different versions of $\mathbf{X}$ considered here do not have much effect on the selections made when computing $\overline{Z}_{vt}$ from MCMC across all loci simultaneously. Very few of the variants that were left out of $\mathbf{X}^{(3)}$ are selected when they are included in a different version of $\mathbf{X}$. In fact, adding correlated variants to $\mathbf{X}$ is more likely to result in a slight reduction in the number of selected variants. In contrast, applying BH to the p-values has a much greater effect on the selections between $\mathbf{X}^{(3)}$ and $\mathbf{X}^{(5)}$, with marginal p-values resulting in many more selections and

Table 9: Cases in which a variant would be selected for one version of $\mathbf{X}$ but not another. Entries show $\overline{Z}_{vt}$; "NA" means that the variant is not in that version of $\mathbf{X}$, while blanks mean that the variant would not be selected for any $\mathbf{X}$.

	none				*Across Traits*				*Across Sites*			
max correlation	0.3	0.5	0.7	0.9	0.3	0.5	0.7	0.9	0.3	0.5	0.7	0.9
HDL												
rs2066715	0.92	0.89	0.23	0.27	0.78	0.75	0.17	0.22	0.92	0.89	0.13	0.13
rs2853579	NA	NA	0.75	0.70	NA	NA	0.75	0.69	NA	NA	0.88	0.87
rs140547417									NA	0.73	0.74	0.66
rs2575875	0.92	0.73	0.40	0.40	0.79	0.66	0.39	0.41	0.94	0.82	0.46	0.47
rs12314392									0.73	0.54	0.53	0.19
rs7946766	0.99	0.97	0.97	0.56	0.98	0.97	0.97	0.52	1.00	0.98	0.98	0.65
rs174546					0.72	0.70	0.62	0.58				
LDL												
rs651821					always >0.99				0.73	0.66	0.63	0.64
TG												
v_c9_107544165	NA	0.86	0.85	0.83	NA	0.83	0.82	0.79	NA	0.87	0.87	0.87
rs174546					0.74	0.74	0.68	0.64				

p-values from the full model resulting in many fewer selections.

Variants with sufficiently strong correlations could, however, lead to convergence problems in our MCMC. To allow arbitrary correlations, we also compute the posterior distribution exactly instead of using MCMC. To achieve this, we make the following simplifications as in Servin and Stephens (2007); Hormozdiari *et al.* (2014); Chen *et al.* (2015):

- Specify τ. We choose $\tau = 0.05$ because the mean of the sampled values of τ in the MCMC was always between 0.04 and 0.06 for all versions of $\mathbf{X}$ with the actual phenotype data.
- Assume the posterior density is zero if $|\mathbf{Z}| \geq K$, where K is typically no more than six but may be as small as one.

Even requiring $|\mathbf{Z}| \leq 4$, however, would still leave far too many subsets if we were to consider all 1302 variants simultaneously, so we only consider one locus at a time. In fact, we consider only two loci: ABCA1 and CETP, both with trait HDL, because these show stronger evidence of influencing the trait via multiple variants. The locus ABCA1 has 58 variants, so we compute the posterior for $|\mathbf{Z}| \leq 6$; whereas the locus CETP has 98 variants, so we require $|\mathbf{Z}| \leq 5$. Finally, we only tried this with our basic model.

In addition to estimating the marginal posterior expected values as $\mathbb{E}[Z_v|\mathbf{y}] \approx \sum_{\mathbf{Z}:Z_v=1} f_{\mathbf{Z},\tau}(\mathbf{Z}, \tau = 0.05|\mathbf{y})$, we also consider two other approaches to selection. One is simply to choose $\mathbf{Z}$ with the largest value of $f_{\mathbf{Z},\tau}(\mathbf{Z}, \tau = 0.05|\mathbf{y})$. The other is to use the confidence set proposed by Hormozdiari *et al.* (2014), described as follows. The probability that a set $\mathcal{S}$ of variants contains all causal variants is $\mathcal{P}(\mathcal{S}) \equiv \sum_{\mathbf{Z}:Z_v=0 \text{ for } v\in\mathcal{S}} f_{\mathbf{Z},\tau}(\mathbf{Z}, \tau = 0.05|\mathbf{y})$. Ideally, one would choose the smallest $\mathcal{S}$ such that $\mathcal{P}(\mathcal{S})$ is above a specified threshold, but this is not computationally feasible. Instead, we use the stepwise selection process in Hormozdiari *et al.* (2014). At each step, the variant that increases $\mathcal{P}(\mathcal{S})$ the most is added to $\mathcal{S}$. (Hormozdiari *et al.* 2014) use the stopping criterion $\mathcal{P}(\mathcal{S}) \geq 0.95$, but this results in a very large number of selections for our data when the maximum value of $|\mathbf{Z}|$ is four or greater. Instead, we use threshold 0.7.

The 70%-confidence set contains all the variants selected by any of the Bayesian methods, so Table 10 lists the variants in the order in which they enter the 70%-confidence set. The selections from BH applied to the p-values for only the one locus and the HDL are also shown, although BH applied to the marginal p-values for locus CETP selects nine additional variants that are not shown. Chen *et al.* (2015) give a good argument for preferring selections based on marginal posterior expectations over the confidence set if there is more than one causal variant, so we do not discuss the confidence sets further here. BH applied to marginal p-values is probably also selecting too many variants. The other methods highlight many of the same variants. Overall, for 90% of the variants in $\mathbf{X}^{(9)}$, the value of $\overline{Z}_{vt}$ computed by MCMC is within 0.01 of the exact posterior marginals without pruning, and the latter value is almost always less than the former, which again suggests that pruning did not result in many missed discoveries. One of the largest differences between these two values is for *rs2853579*, which—as noted above in connection with Table 9—is strongly correlated with *rs2066715*. Furthermore, for locus ABCA1, the set $\mathbf{Z}$ of variants with the largest posterior density is equal to the variants selected by using $\overline{Z}_{vt}$ with $\mathbf{X}^{(3)}$. For the locus CETP, these two sets cannot be the same because $\overline{Z}_{vt}$ selects six variants, while we had to assume $|\mathbf{Z}| \leq 5$ in order to compute the exact distribution; additionally, the mode of the posterior density includes one very rare variant (MAF 0.0002) that is in $\mathbf{X}^{(3)}$ but that the basic prior does not select.

For completeness, the detailed descriptions of the differences in Table 9 are as follows:

rs2066715 **and** *rs2853579***,** which are both in ABCA1, are in the table for all three priors applied to response HDL. Their correlation is just below 0.7, and $\overline{Z}_{vt}$ for the former decreases drastically when the latter is in $\mathbf{X}$ and is selected instead—although $\overline{Z}_{vt}$ for *rs2853579* falls just below the selection threshold with *Across Traits* applied to $\mathbf{X}^{(9)}$. The former is a missense variant predicted to be benign; whereas the latter is a synonymous variant.

rs140547417 is not in $\mathbf{X}^{(3)}$, has $\overline{Z}_{vt}$ barely above the selection threshold for $\mathbf{X}^{(5)}$ and $\mathbf{X}^{(7)}$ when *Across Sites* is applied to HDL, then drops slightly below the threshold for $\mathbf{X}^{(9)}$. Its minor allele occurs in only three subjects, one of which has the 32nd largest value of HDL and another is in the 15th-percentile. Furthermore, this variant is in gene CETP, and its group includes a variant with $\overline{Z}_{vt} \approx 0.8$ and one with $\overline{Z}_{vt} \approx 0.3$.

rs2575875 is strongly correlated with a new variant in $\mathbf{X}^{(5)}$ (with $\overline{Z}_{vt} = 0.24$) and with another new variant in $\mathbf{X}^{(7)}$—at which point all three variants have $\overline{Z}_{vt}$ in 0.3–0.4.

rs12314392 has $\overline{Z}_{vt}$ barely above the selection threshold when *Across Sites* is applied to $\mathbf{X}^{(3)}$. It has correlation 0.3–0.4 with three variants that are in $\mathbf{X}^{(5)}$. Two of these variants have $\overline{Z}_{vt} < 0.05$, while the other has $\overline{Z}_{vt} = 0.2$; this is easily enough for *rs12314392* to drop below the selection threshold. None of the new variants in $\mathbf{X}^{(7)}$ are strongly correlated with *rs12314392*, but four of the new variants in $\mathbf{X}^{(9)}$ have correlations 0.43–0.84 with *rs12314392*. Three of these have $\overline{Z}_{vt} < 0.1$; but the one with correlation 0.75 has $\overline{Z}_{vt} = 0.60$, causing $\overline{Z}_{vt}$ for *rs12314392* to decrease even more.

rs7946766 has $\overline{Z}_{vt} > 0.96$ for HDL with all three priors until several variants strongly correlated with it are added to $\mathbf{X}^{(9)}$, including two that have $\overline{Z}_{vt} \approx 0.3$.

v_c9_107544165 has minor allele that occurs in only two subjects, one of which has the 2nd largest value of HDL and the other is in the 14th-percentile. It is in the UTR 3′ of ABCA1, which means it is in a group by itself in the *Across Sites* prior.

Table 10: Comparison of selection methods to see effect of using all variants. These tables list all variants in the 70%-confidence set in the order in which they would be added to this set. The variant names in bold face have $\overline{Z}_{vt} > 0.7$ from the MCMC with $\mathbf{X}^{(3)}$, while those preceded by an asterisk are *not* in $\mathbf{X}^{(3)}$. In the columns for $\overline{Z}_{vt}$ from MCMC, a blank means that $\overline{Z}_{vt} < 0.2$, and NA means that the variant is not in that $\mathbf{X}^{(9)}$. In the columns for the exact marginal, a blank means that the value is less than 0.2.

locus ABCA1 with response HDL

	$\mathbf{X}^{(9)}$	all variants in locus			
variant	MCMC $\overline{Z}_{vt}$	exact marginal	argmax$_{\mathbf{Z}}$ $f(\mathbf{Z}\|\tau,\mathbf{y})$	BH full	BH marginal
rs2575875	0.40	0.48	yes		yes
v_c9_107555452	0.99	0.97	yes	yes	yes
v_c9_107548661	0.85	0.81	yes	yes	yes
rs2066718	0.93	0.73	yes	yes	yes
rs2066715	0.27	0.28	yes		yes
v_c9_107555091	0.87	0.62	yes	yes	yes
*rs1800978	0.33	0.26			yes
*rs2853579	0.70	0.32			yes
*rs2740486	0.36	0.25			yes
rs76881554	0.46				
*rs2066714	NA				yes
*rs2230806					
*rs2066716					yes
end of 50%-confidence set					
*rs2230805	NA				
rs2230808	0.24				
*v_c9_107544165					
rs41437944					
rs145183203					
rs9282537					

locus CETP with response HDL

	$\mathbf{X}^{(9)}$	all variants in locus			
variant	MCMC $\overline{Z}_{vt}$	exact marginal	argmax$_{\mathbf{Z}}$ $f(\mathbf{Z}\|\tau,\mathbf{y})$	BH full	BH marginal
rs3764261	1.00	1.00	yes	yes	yes
rs5883	1.00	0.98	yes	yes	yes
rs5880	0.95	0.92	yes		yes
rs34216426	0.86	0.62	yes	yes	yes
v_c16_57095439	0.46	0.28	yes	yes	yes
rs2303790	0.92	0.29		yes	yes
rs5801	0.86	0.26			yes
end of 50%-confidence set					
*v_c16_56926936					
rs140440847					yes
*rs140547417					
v_c16_57117215	0.22				
*rs12445698					yes
rs3751705				yes	

rs174546 **and** *rs651821* both have $\overline{Z}_{vt}$ between 0.7 and 0.75 for $\mathbf{X}^{(3)}$, but it decreases as more variables are added to $\mathbf{X}$ until it is between 0.58 and 0.65 for $\mathbf{X}^{(9)}$; in other words, these variants are near the threshold for selection for all versions of $\mathbf{X}$.